1974-06

A time-domain waveform display/analyzer for speech research

Edmondson, Gary Duane

Monterey, California. Naval Postgraduate School

A TIME-DOMAIN WAVEFORM DISPLAY/ANALYZER FOR SPEECH RESEARCH

Gary Duane Edmondson

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

A Time-Domain Waveform
Display/Analyzer for Speech Research

by

Gary Duane Edmondson

June 1974

Thesis Advisor: G. D. Ewing

Approved for public release; distribution unlimited.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Time-Domain Waveform Display/Analyzer for Speech Research		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis June 1974
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Gary Duane Edmondson		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		12. REPORT DATE June 1974
		13. NUMBER OF PAGES 91
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Naval Postgraduate School Monterey, California 93940		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

visible speech
speech analysis
speech recognition
pattern recognition
waveform analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A system has been developed for visually displaying the up-crossings (zero-crossing with positive slope) of a speech acoustic waveform on a storage CRT. Studies of these displays have revealed readily distinguishable visual patterns which are useful for discriminating some consonants, often even in connected speech. Representative displays are included to illustrate the salient features of this visual speech display. Based on preliminary investigations, the system shows promise for application as a speech training aid for deaf children and

Item 20 (continued)

as a cross-cultural language training aid. Complete design data and schematics for the visual Time-Domain Display/Analyzer are included with recommendations for utilizing the system in evaluating time-domain parameter extractors based on zero-crossings in future speech research. This Time-Domain Waveform Display/Analyzer system provides valuable results for comparing parameter extractors based on zero-crossings and indicates how they might be related iteratively or combined to yield better extractors.

A Time-Domain Waveform

Display/Analyzer for Speech Research

by

Gary Duane Edmondson
Lieutenant, United States Navy
B.S., Univeristy of New Mexico, 1966
M.S., Naval Postgraduate School, 1973

Submitted in partial fulfillment of the
requirements for the degree of

ELECTRICAL ENGINEER

from the

NAVAL POSTGRADUATE SCHOOL
June 1974

ABSTRACT

A system has been developed for visually displaying the up-crossings (zero-crossing with positive slope) of a speech acoustic waveform on a storage CRT. Studies of these displays have revealed readily distinguishable visual patterns which are useful for discriminating some consonants, often even in connected speech. Representative displays are included to illustrate the salient features of this visual speech display. Based on preliminary investigations, the system shows promise for application as a speech training aid for deaf children and as a cross-cultural language training aid. Complete design data and schematics for the visual Time-Domain Display/Analyzer are included with recommendations for utilizing the system in evaluating time-domain parameter extractors based on zero-crossings in future speech research. This Time-Domain Waveform Display/Analyzer system provides valuable results for comparing parameter extractors based on zero-crossings and indicates how they might be related iteratively or combined to yield better extractors.

TABLE OF CONTENTS

I. INTRODUCTION

After many centuries of sporadic interest in the nature of Speech, the past twenty years of speech research stand out as particularly intensive. Much time, energy, and expense have been expended in studying the acoustics of speech both for basic research and for applications in speech bandwidth compression, linguistic studies, automatic speaker identification, speech aids for the handicapped, child speech development, cross-cultural language translators, and speech recognition.

A. OBJECTIVES

One of the principle objectives of this thesis was to design and implement an inexpensive speech waveform analyzer for research and application in human speech recognition. The particular application in speech recognition is in the areas of speech aids for the handicapped and child speech development. The ultimate goal was to develop a system inexpensive enough to be in economic reach of the individual handicapped person and small institutes or clinics with limited financial resources.

B. RECENT TREND IN SPEECH RECOGNITION RESEARCH

In the 15 years, between 1950 and 1965, there had been developed a number of electronic machines that would recognize very limited vocabularies pronounced by particular speakers for whom the machines had been adjusted. Lindgren [1] , in an interesting survey of machine recognition

of speech, follows the development of the automatic speech-recognition art from the simple analog hardware systems of the early 1950's to the more sophisticated computer software systems of the mid 1960's. Since Lindgren's informative survey, the application of computer based analysis has greatly intensified. Recent literature is replete with an ever growing volume of data generated by the power of the modern computer. However, in keeping with the ultimate goal of this thesis, it quickly became apparent that analysis by computer, even the most inexpensive "bare-bones" mini-computer available, was too expensive for serious consideration at this time for the stated objective.

C. FREQUENCY-DOMAIN VS TIME-DOMAIN ANALYSIS

For the purposes of this thesis, frequency-domain analysis techniques are defined as all fourier transform processes which extract spectral components from the speech waveform. On the other hand, time-domain analysis techniques attempt to extract the significant features of speech from the time waveform rather than from the frequency spectrum. Specific examples of time-domain analysis techniques are zero-crossings, auto-correlation, and linear prediction of the time waveform.

A moderately comprehensive review of the literature of acoustic speech analysis over the past 25 years will quickly convince one that the majority of this work has been conducted through the application of spectrographic and other frequency-domain studies of the speech acoustic waveform. Speech waveforms are voltage-time plots directly proportional to the changes in air pressure produced in the speech process. Such

voltage-time plots or oscillograms contain complete information about the original signal from which it was derived. An oscillogram is an analog time-domain display from which the precise time of occurrance of an event or feature, such as the time of maximum amplitude, dV/dt, zero-crossing, etc., may be determined. Generally the speech waveform exhibits considerable short-time variation and a given acoustic state quickly transitions into a very different one. Unfortunately past studies have not revealed efficient techniques for reducing this large bulk of highly variable data and deriving from it useful speech parameter, particularly in real-time. Therefore, until the advent of modern computer processing, time-domain analysis of the waveform itself was relegated to a very minor role in speech research; the major role being played by frequency-domain analysis.

There are several important differences between direct time-domain and classical frequency-domain analysis techniques. To carry out a frequency-domain/analysis (spectrographic study), it is necessary to use a large number of contiguous bandpass filters implemented in analog or digital form. Such a spectral analysis averages waveform frequency components over a period of time fixed by the uncertainty principle, i.e., the law governing the relationship between the bandwidth of the spectral component filter and the resolving time. This relationship may be stated by the following equation:

$$T_r \geq \frac{1}{W} \qquad (1)$$

where T_r is the filter response time and W is the bandwidth, in Hz, of the filter associated with the spectral component resolved.

In consonance with the above principle, spectral analysis of steady state periodic or quasi-periodic signals provide accurate frequency components of the waveforms. In speech, only vowel phonemes satisfy this criteria and are therefore most amenable to this analysis technique. In vowel phonemes (voiced speech), the vocal cords vibrate between open and closed positions releasing quasi-periodic pulses of air. The fundamental frequency of this vibration is termed pitch. The air pulses, in turn, excite the vocal tract where the positions of speech articulators, such as the tongue and palate, produce resonance conditions. These resonance conditions concentrate the acoustic energy into specific areas of the frequency spectrum, known as formants.[1] The effect of both periodic excitation and oral cavity configuration are evident in the waveform structure of voiced speech. The waveform repeats periodically at the pitch-frequency interval; whereas between each such interval, its structure represents the dominant formants.

In consonant phonemes (unvoiced speech), the vocal cords do not vibrate. Instead, air turbulence resulting from either the passage of air through a narrow constriction formed by the articulators or the sudden release of air (stop consonants) by the lips or tongue, creates acoustic noise that excites the vocal tract. As in voiced speech, the articulators create resonance conditions that concentrate the unvoiced acoustic energy into particular areas of the frequency-power spectrum. The spectral

energy of unvoiced speech is continuous with frequency, in constrast to voiced speech where the energy occurs as discrete frequency components.

In semi-vowel phonemes, the combined effects of voiced and unvoiced speech are evident as a noise-like spectrum with a super-imposed resonance structure characteristic of vocal cord activity.

The consonant phonemes, with their non-periodic, noise-like waveforms, often undergo rapid frequency and amplitude modulations and are not adequately described by frequency-domain analysis. Here the bandwidth limitation becomes a serious hindrance because the time during which a feature of a consonant phoneme occurs is often too short for accurate frequency discrimination.

The implications of frequency averaging imposed by the uncertainty principle must be carefully considered. Assume that acoustic invariants (specific, measurable features that are invariant among a population of speakers) are embedded in a given phoneme waveform, as surely they are when humans recognize nonsense syllables for example. Further assume that, as common for most consonants, this phonome is very irregular in that it has rapidly changing frequency characteristics. When classical spectral analysis methods are applied to this signal, the result is an average of these frequencies and does not bear a unique relationship to the original signal. In fact, the magnitude spectrum of a given waveform over the interval of averaging is the same whether the waveform is played backward or forward. And, clearly, averaging over a period of changing frequency does not yield a good representation of the original signal.

Due to the redundancy in speech, analysis in the time-domain yields information found by spectral analysis and vice versa. That is, time-domain analysis reveal some information about the steady state vowels, just as frequency-domain analysis reveal some information about the transitional state consonants.

D. RESULTS OF FEASIBILITY STUDY

With the above considerations in mind, further research was undertaken to determine an approach which would achieve results subject to the constraints of low cost and small size while providing a moderate degree of automatic data processing.

A study of spectral analysis techniques revealed little insight into methods of extracting the required spectral information differing from previous filter bank or Fast Fourier Transform (FFT) implementations which had already been determined to be excessively costly.

Turning then to time-domain analysis, there appeared to have been less formalized approaches to the problem and more ad hoc experimentation except in the areas concerning zero-crossing averaging to obtain spectral information and in recent computer-aided auto-correlation and linear prediction techniques. A careful review of the meager literature revealed a connective link between several of the more successful ad hoc systems which is explained in the recommendation section of this thesis. After some preliminary evaluation, it appeared that this connective link might best be investigated by utilizing the inherent pattern recognition capability of the human observer of a time-domain visual display.

The display portion of the analyzer has been constructed with extensive utilization of analog and digital integrated circuits to obtain the desired result while reducing circuit complexity and cost. Special processing techniques have been incorporated into the basic display portion of the analyzer to allow expansion of the system to investigate the more desirable features of previously developed recognizers as well as allowing automatic statistical analysis of the time-domain waveform.

E. BASIS FOR THE SYSTEM INSTRUMENTED IN THIS THESIS

The particular time-domain analyzer technique incorporated in the time-domain waveform display/analyzer instrumented in this thesis is based on the studies 25 years ago by Licklider and his colleagues [2,3], who demonstrated the intelligibility of infinitely clipped speech, This showed that a major portion of speech information is encoded in the zero-crossings of the waveform, although this information is probably redundantly present in other waveform features. Since Licklider's studies, other investigators [References 4 - 12] have looked at zero-crossings and up-crossings (zero-crossings with positive slope) but have generally averaged these crossings in time, thereby losing the perfect time resolution available from the unaveraged data. Generally their intent was to find an inexpensive method to extract classical frequency-domain acoustic features: e.g., formants, etc., from averaged time-domain parameters. This would require much less computation than demanded by Fast Fourier Transform techniques and much less hardware than filter-bank techniques. However, the time-averaged approximations to these acoustic features were usually more

variable and less reliable than those same features extracted directly with frequency-domain analysis. By averaging, much of the inherent value of direct time-domain analysis was lost, specifically the opportunity of observing significant, though very short duration, acoustic features.

II. SYSTEM DESCRIPTION

A. INTRODUCTION: AN OVERVIEW

Visual display patterns emerge only after interpretation by the mind. It is necessary to ignore some features and to concentrate on others to bring a pattern forward. The search for patterns is a natural function of the mind and precisely involves the ability to choose, to exclude, to exaggerate, and to minimize. The mind would be expected to function this way in interpreting visual patterns as it does in interpreting acoustical patterns.

To enable the eye-brain system to obtain the information in a speech ocillogram, some form of transformation must be made. Numerous devices to accomplish this transformation have been designed and used, each usually with a specific purpose in mind. These devices range from the familiar "sound spectrograph" to scanned filter banks each performing a time-domain to frequency-domain transformation. [References 13 - 20] . However, to display the information coded in the zero-crossings of an infinitely clipped speech waveform requires a transformation that conforms to the criteria set forth below.

To be successful, a visual display should encourage the eye-brain perceptual process to proceed from gross overall pattern classification, to localized feature recognition, and finally to specific time-evolutions of features. One time-domain display technique described in the literature [5] offered a relatively simple display by channelizing the up-crossing

interval into 14 variable width channels. It did not, however, meet the perceptual criterion as set forth above. Direct display of up-crossing time-interval vs real-time was unsatisfactory because the inverse time-frequency relationship compressed the up-crossings produced by consonant phonemes into only one-tenth of the display. Reciprocal time-interval vs real-time displays linerized the display but were far from being perceptually acceptable.

The perceptually acceptable time-domain display that evolved from this investigation is conceptually similar to one described by Baker [21]. The display is generated as follows:

Analog circuitry was designed to generate a pulse at the exact time the input speech voltage waveform crosses a preset threshold voltage in a positive going direction as illustrated by waveform B in Figure 1. The preset threshold is set slightly above the peak amplitude of the background noise to provide a degree of noise rejection during periods of silence. The intervals between successive pulses (up-crossings) are transformed and displayed as a function of time or as a function of total up-crossings occurring in an utterance. The transformation of successive up-crossing intervals consists of generating a voltage proportional to the Logarithm of the Reciprocal of the Up-Crossing Interval (Log RCI). This parameter (LogRCI) is applied to the vertical input of a storage oscilloscope and the z-axis is strobed ON for 10 μs following each up-crossing interval. Therefore, each up-crossing interval in the acoustic waveform is represented in the CRT display by a discrete intensified dot whose vertical position is the LogRCI parameter and whose horizontal position is made

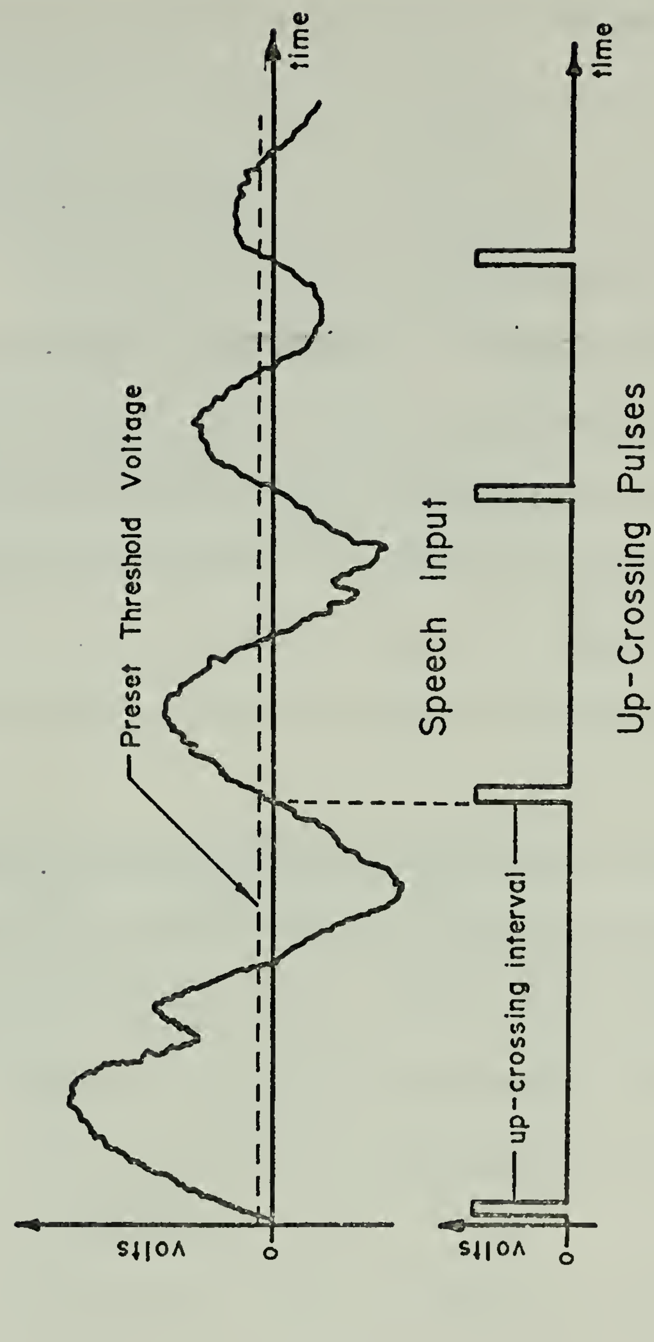

time
Preset Threshold Voltage
Speech Input
Up-Crossing Pulses
up-crossing interval
volts
0
time
volts
0

Figure 1.

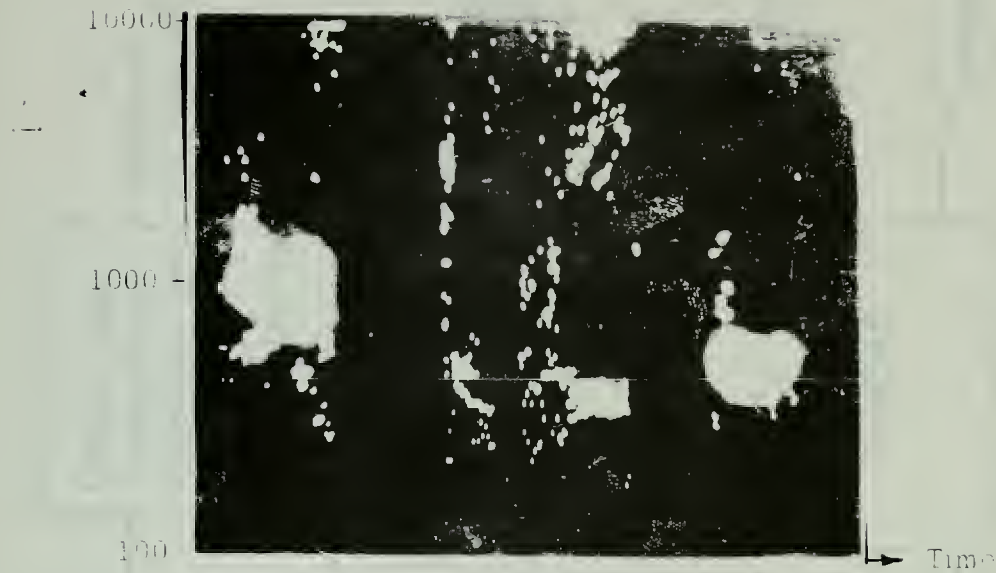
proportional to the elapsed time since the start of the utterance or to the total up-crossings which have occurred since the start of the utterance. This yields a display that superficially resembles a sound spectrograph as illustrated in Figure 2.

B. BASIC THEORY OF OPERATION

A block diagram of the display section of the analyzer is shown in Figure 3. The speech signal is obtained from a microphone indicated in the upper left of the diagram or from a high level tape recorder input. This signal is amplified and dynamically amplitude limited by a fast-acting, low distortion AGC/Squelch Amplifier. This increases the microphone signal to a useable level and compensates for talker loudness variations and for the difference in loudness of the voiced vs the unvoiced portions of speech.

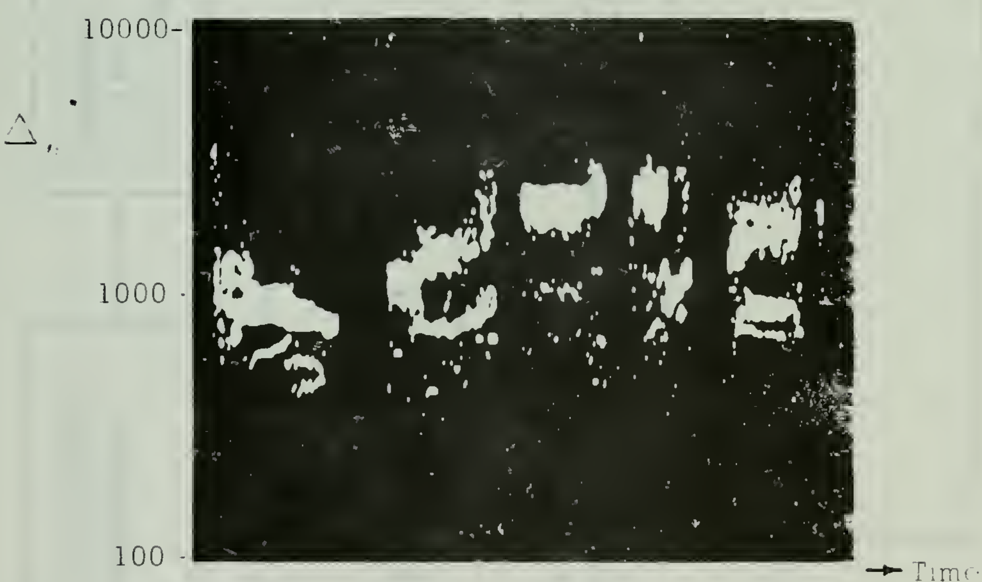
The speech signal is then band-limited from 100Hz to 10KHz before it is applied to a Logarithmic Threshold-Crossing Detector (LTD). The LTD produces a 25 microsecond TTL compatible positive pulse (waveform B of Figure 4) when the input signal crosses a preset threshold voltage going in a positive direction. For a zero voltage threshold, the detector exhibits a 50 DB (-35 dbm to +15 dbm) dynamic range.

The Timing Control Generator (TCG) requires TTL logic level inputs from the AGC/Squelch Amplifier, the Logarithmic Threshold-Crossing Detector, and the Log RCI Generator. If the voice-operated switch (VOX) (derived from the squelch detector of the AGC/Squelch Amplifier) is ON, the Timing Control Generator produces a sequence of pulses following each



The phrase "Pawn to queen four"
by a male speaker (HR). HIGH
Threshold setting.

*Arbitrary parameter defined by equation 15



The phrase "Oh my aching back"
by a male speaker (LE). LOW
Threshold setting.

Figure 2

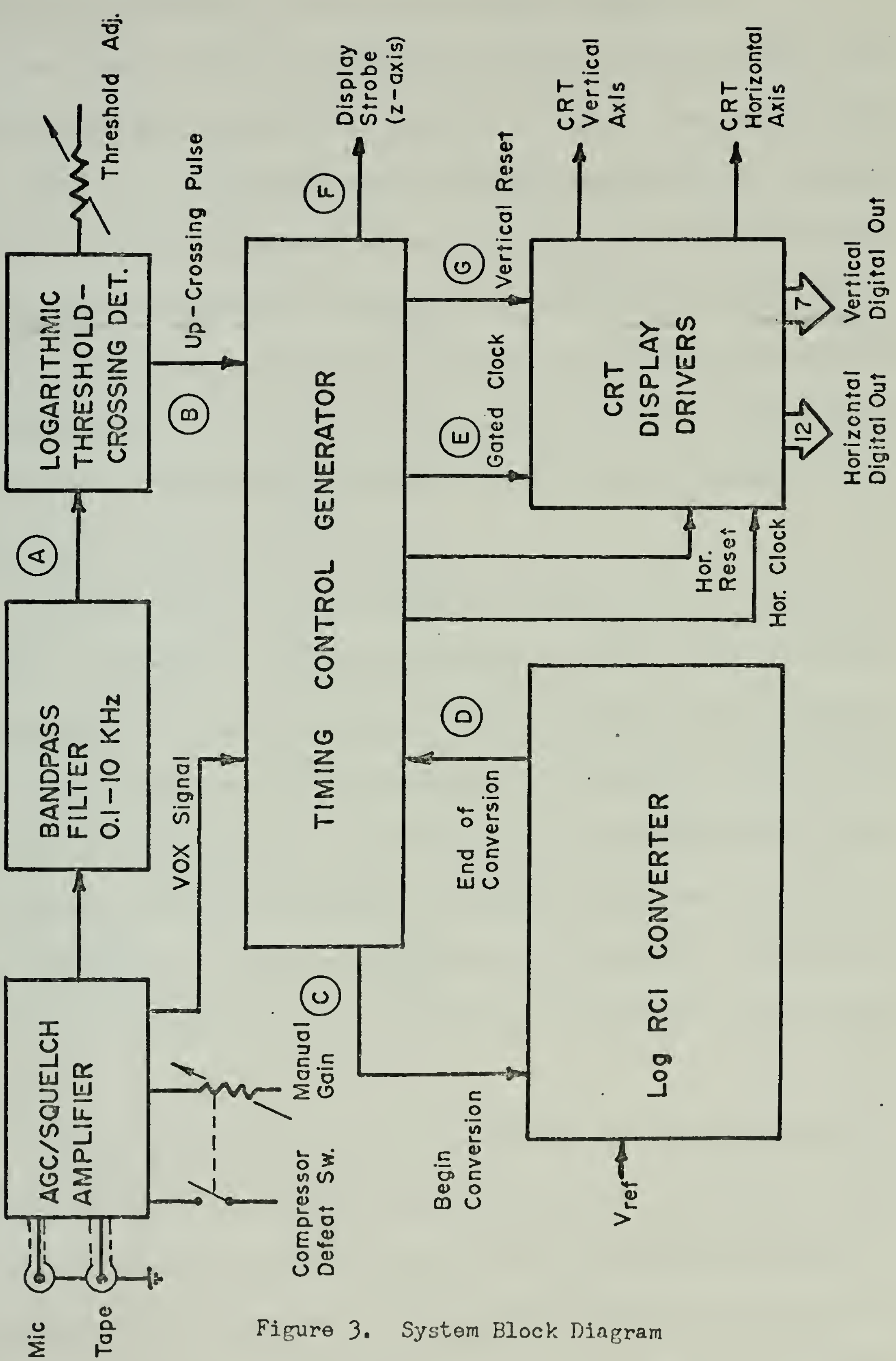


Figure 3. System Block Diagram

up-crossing pulse input. These pulses control the transformation of the proceeding up-crossing interval and the display presentation.

The timing waveforms are depicted in Figure 4 and are labeled with corresponding letters in the block diagram of Figure 3. The time interval between the low-to-high transition of the Begin Conversion Pulse (waveform C) to the high-to-low transition of the End Conversion Pulse (waveform D) is logarithmicly related to the immediately preceding up-crossing interval. This conversion time interval varies from 300 nanoseconds to 64 microseconds corresponding to a 100 microsecond and a 10 millisecond up-crossing interval respectively. Up-crossing intervals of shorter or greater duration are inhibited by the Timing Control Generator.

The conversion time interval gates a 2.0 MHz clock (waveform E) to the Vertical Display Driver which consists of a CMOS seven-bit counter D/A converter. The analog voltage output (waveform H of Figure 7b) from the Vertical Display Driver is proportional to the Logarithm of the Reciprocal of the Up-Crossing Interval (Log RCI). Seventy microseconds after each up-crossing interval a 10 microsecond Display Strobe (waveform F) is generated that turns on the beam intensity of a storage CRT producing the dot display. Waveform G is the reset pulse for the Vertical Display Driver counter.

The Horizontal Display Driver is a CMOS twelve-bit counter/DA converter driven by an operator adjustable oscillator in the Timing Control Generator or by the up-crossing pulses from the Logarithmic Threshold-Crossing Detector. The counter is reset automatically when the VOX turns off or it may be manually reset by the operator. The output of this driver

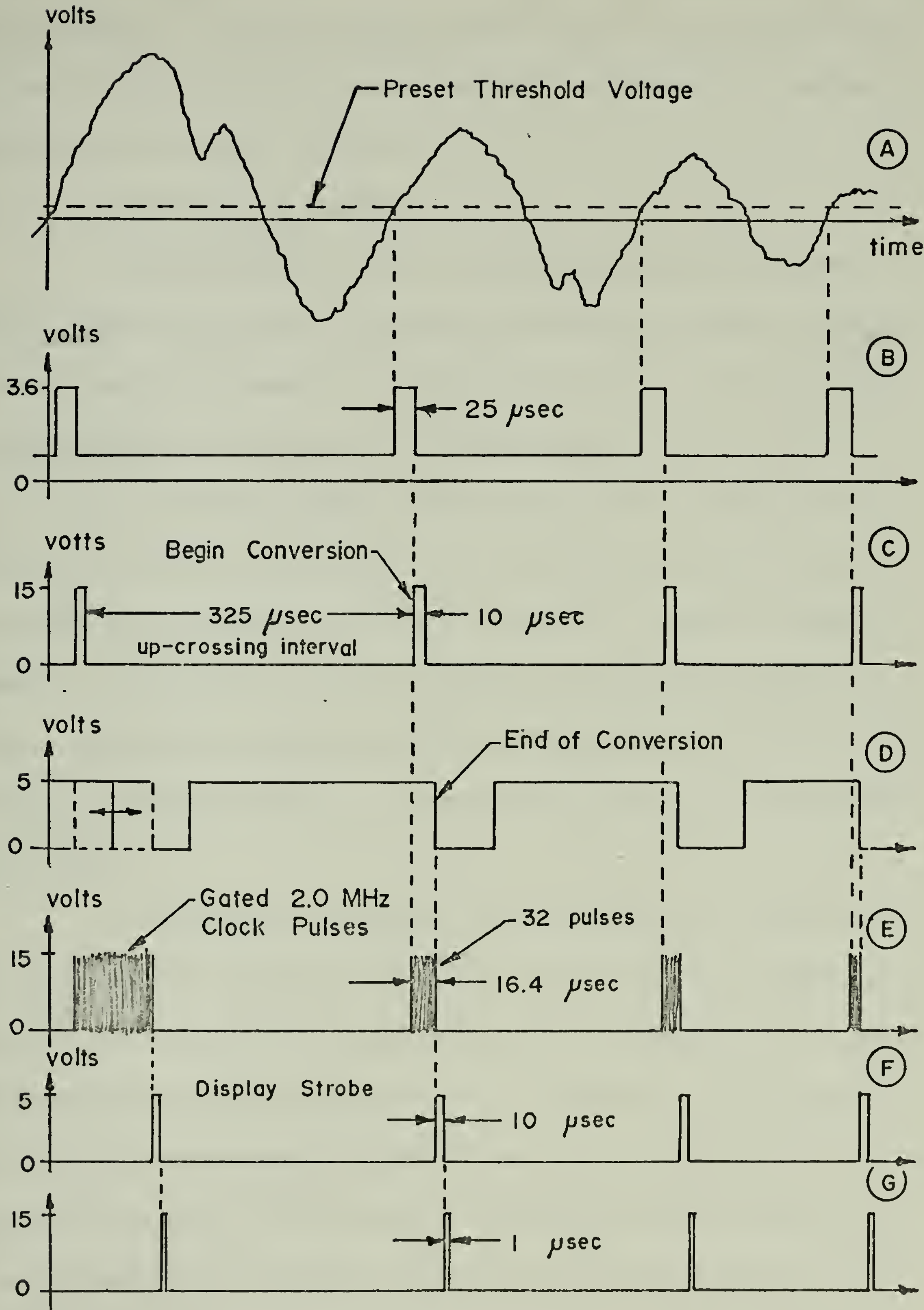


Figure 4. System Timing Diagram

is a linearly increasing staircase with real-time or accumulated up-crossings as the parameter. The digital output of this counter is also called out for external circuitry, thereby allowing real-time and relative time coding of feature occurrence in the display.

1. <u>AGC/Squelch Amplifier</u>

A monolithic AGC/Squelch amplifier (A1) shown in Figure 5, with its associated circuitry, provides low distortion, dynamic amplitude compression over a signal input range exceeding 70 Db. The monolithic realization of the amplifier with its characteristics are described in Ref. 41. This low cost IC is ideally suited to applications requiring low distortion audio AGC with fast-attach/slow-release characteristics while providing 40 db of signal voltage gain. In addition, a built-in squelch detector, which takes its input before the AGC stage, may be used as a fast-attack voice operated switch (VOX) with adjustable switching threshold. Reliable switching action may be obtained down to below one millivolt of signal input.

In summary, the essential features of this input processing circuit are to (1) control the peak amplitude of the speech output signal within $\pm$ 1.0 Db over an input signal variation of greater than 70 Db with fast-attack/slow-release characteristics, (2) provide a voice operated switch (VOX) function which exhibits a high degree of discrimination between silence and speech in the presence of moderately low level background noise to preclude the necessity of using a "quiet room" (although a high quality noise-canceling microphone is desirable), and (3) provide the VOX

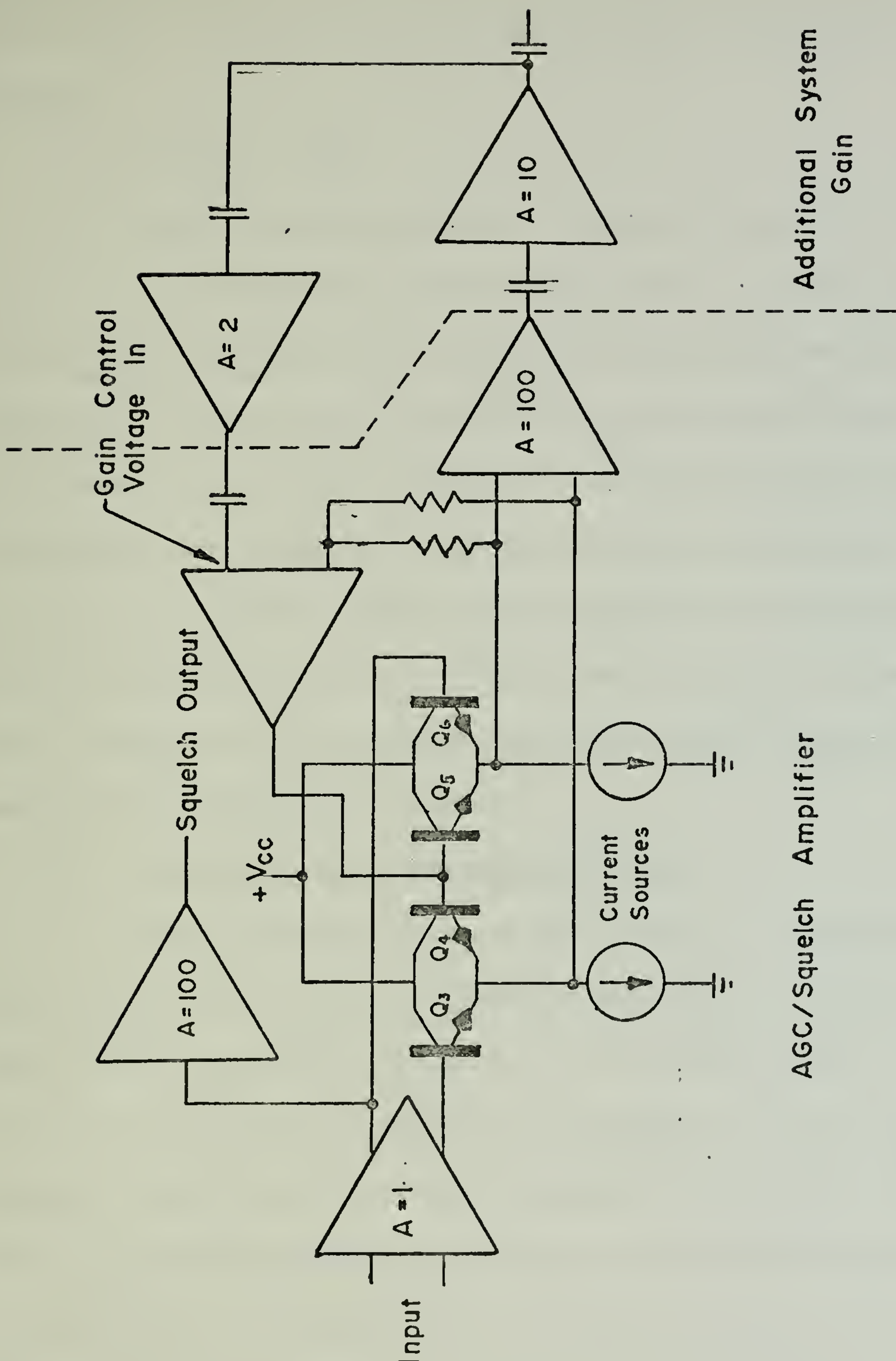


Figure 5. AGC/Squelch Amplifier

signal in the first up-crossing interval of speech so that no initial information is lost. A more detailed circuit diagram may be found in the Appendix.

2. Band-Pass Filter

Amplifier A4 and A5 of Figure 31 (Appendix) constitute 2-pole Bessel low-pass and high-pass filters respectively. The low-pass filter (A4) roll-off is 40 db/decade above 10 KHz to remove low level high frequency noise while still retaining the essential high frequency components of unvoiced speech. Low frequency roll-off is 60 db/decade below 100 Hz to exclude power line interference and low level rumble associated with microphone inputs. The output of this filter is brought out to a rear panel tape recorder output jack for recording of band-limited, amplitude compressed speech signals for use in comparative analysis. Further design and filter equations are contained in the appendix.

3. Logarithmic Threshold-Crossing Detector

Although the input signal to this stage is well controlled in peak amplitude, there may be short-time (approximately 20 percent of AGC release time) variations as great as 10 Db. In the manual mode of operation this signal variation will consistently exceed 30 Db for normal speech. Therefore, it was desirable to design a threshold crossing detector that would accept this input range and, in addition, provide TTL compatible logic level outputs. For application versatility several other features were included in the design, such as wide range of threshold adjustment above and below the input zero voltage axis, a defined lower threshold with snap transition for noise free zero-axis detection, and wide bandwidth. The

circuit typically responds, noise free, from -35 Db to +15 Db referred to 0.775 volts. The output of comparator A1 (Figure 32, Appendix) drives the schmitt-trigger (B) input of a 74121 TTL monostable multivibrator which is set to produce a 25 μs pulse at the positive transition of the comparator.

4. Timing Control Generator

The inputs to this circuit are as follows:

(1) Up-crossing transitions (UT) from the Logarithmic Threshold-Crossing Detector.

(2) A voice-operated-switch (VOX) signal from the AGC/Squelch Amplifier.

(3) An End of Conversion (EOC) signal from the Log RCI Converter.

The outputs of this circuit arc as follows: (relative timing and destination may be determined by referring to Figure 4)

(1) A Begin Conversion (BC) signal.

(2) Gated Vertical Display Driver clock pulses (VDCP).

(3) Vertical Display Driver Reset pulse (VDR).

(4) Display Strobe (DS) pulse.

(5) Horizontal Display Driver clock pulses (HDCP).

(6) Horizontal Display Drive reset pulses (HDR).

(7) Integrator Reset Pulse (IR).

A simplified circuit diagram and theory of operation are included in the Appendix.

5. Up-Crossing Interval to Log RCI Conversion

There are several methods for converting a time interval into a digital number or analog voltage that is proportional to the logarithm of the time interval. However, for the system described in this thesis it was essential to have available both a digital and an analog representation of this quantity. Stringent constraints are placed on the conversion circuit due to the real-time or near real-time requirements. The conversion, display, and reset must be complete within 100 μs following each up-crossing pulse for conversion of 10KHz signals.

A survey of logarithmic analog-to-digital techniques is given in Ref. 22. References 23 - 28 show various analog voltage to log voltage converters. The different schemes can be grouped in three main classes according to the block diagrams in Figure 6. [22]. The first scheme (Figure 6a) using the Logarithmic Characteristic of a PN Junction, was either too complex to be economically suitable or too slow to produce real-time output. The digital logarithmic converter (Figure 6b) and the logarithmic A/D converter (Figure 6c) were especially complex, but very fast.

Figure 7a is a simplified diagram of a very satisfactory converter for this application. This scheme combines simplicity with moderate accuracy and very low cost -- the ideal combination. This converter is of the form of Figure 6a but does not use the PN Junction characteristics.

a. Time to Voltage Conversion

Operational amplifier A_1 (Figure 7a) is connected to perform the mathematical operation of integration according to the following equation

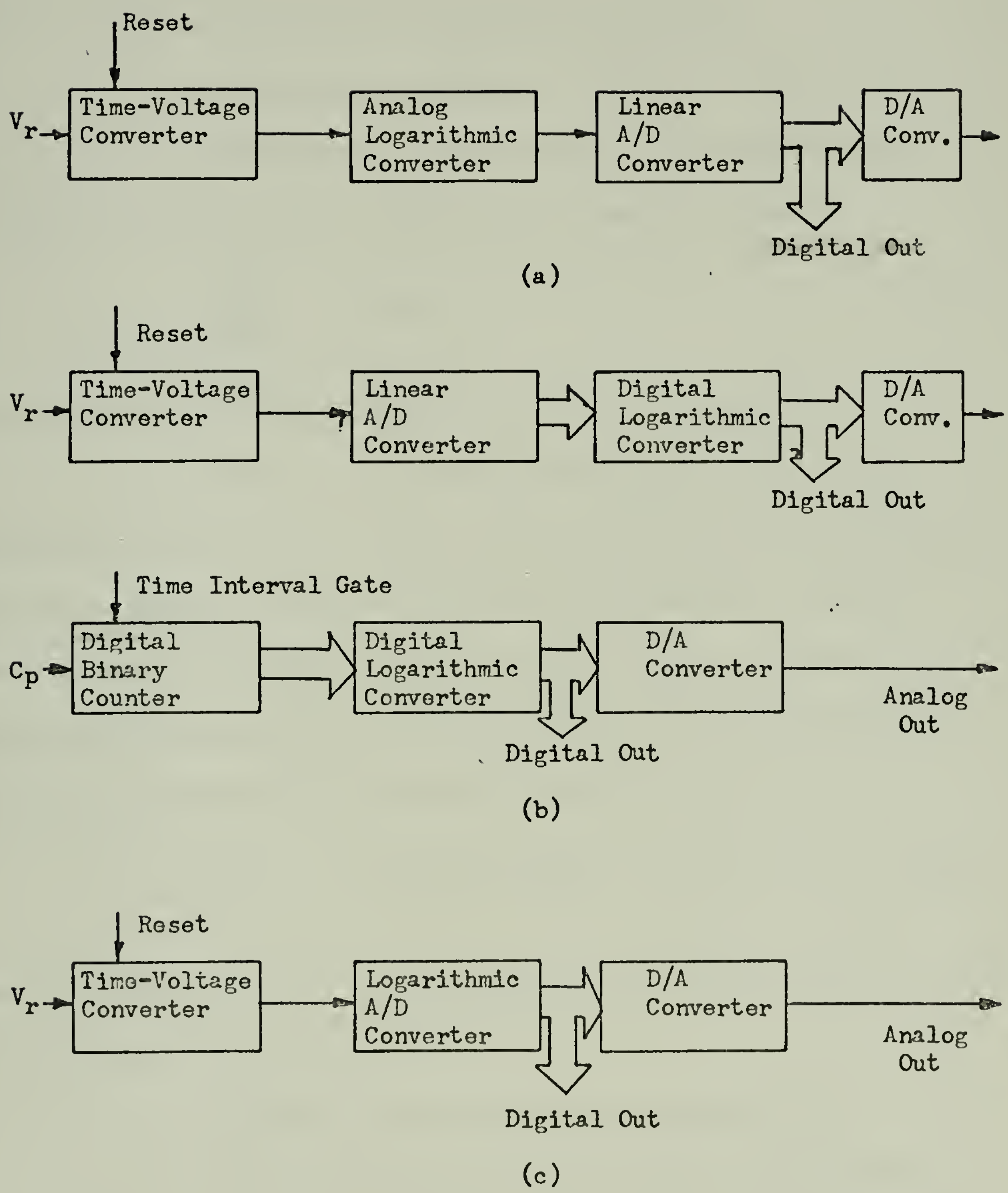


Figure 6. Logarithmic Analog-to-Digital Techniques.

$$v_o = -\frac{1}{C}\int i\,dt = -\frac{1}{RC}\int v_i\,dt \qquad (2)$$

where v_o is the output voltage and v_i is the input voltage.

The amplifier, therefore, provides an output voltage proportional to the integral of the input voltage. If the input voltage is a constant, $v_i = V_r$, then the output will be a ramp described by

$$v_o = -\frac{V_r}{RC}t = -kt \qquad (3)$$

where V_r = -1.0 Vdc ; RC = 1.0 msec ; thus k = 1.0 volt/msec.

Integrator A_1 is reset at the occurrence of each up-crossing pulse (waveform B of Figure 7b) by a CMOS FET switch connected across the integration capacitor. The output is, therefore, a series of sawtooth waveforms whose period, T, is equal to the up-crossing interval. An analog voltage proportional to the up-crossing interval is available at the output at the instant of reset and is given by

$$v_o = -kT_n \qquad n = 1, 2, \cdots, m \qquad (4)$$

where n is the up-crossing interval index number and m is the total number of up-crossing in the utterance.

b. RC Discharge Voltage-to-time Conversion

The output of integrator A_1 is applied to a modified track-hold circuit composed of operational amplifier A_2 (Figure 7a). The modification to the track-hold circuit consists of connecting a resistor across the hold capacitor. The instantaneous voltage, kt, across the hold

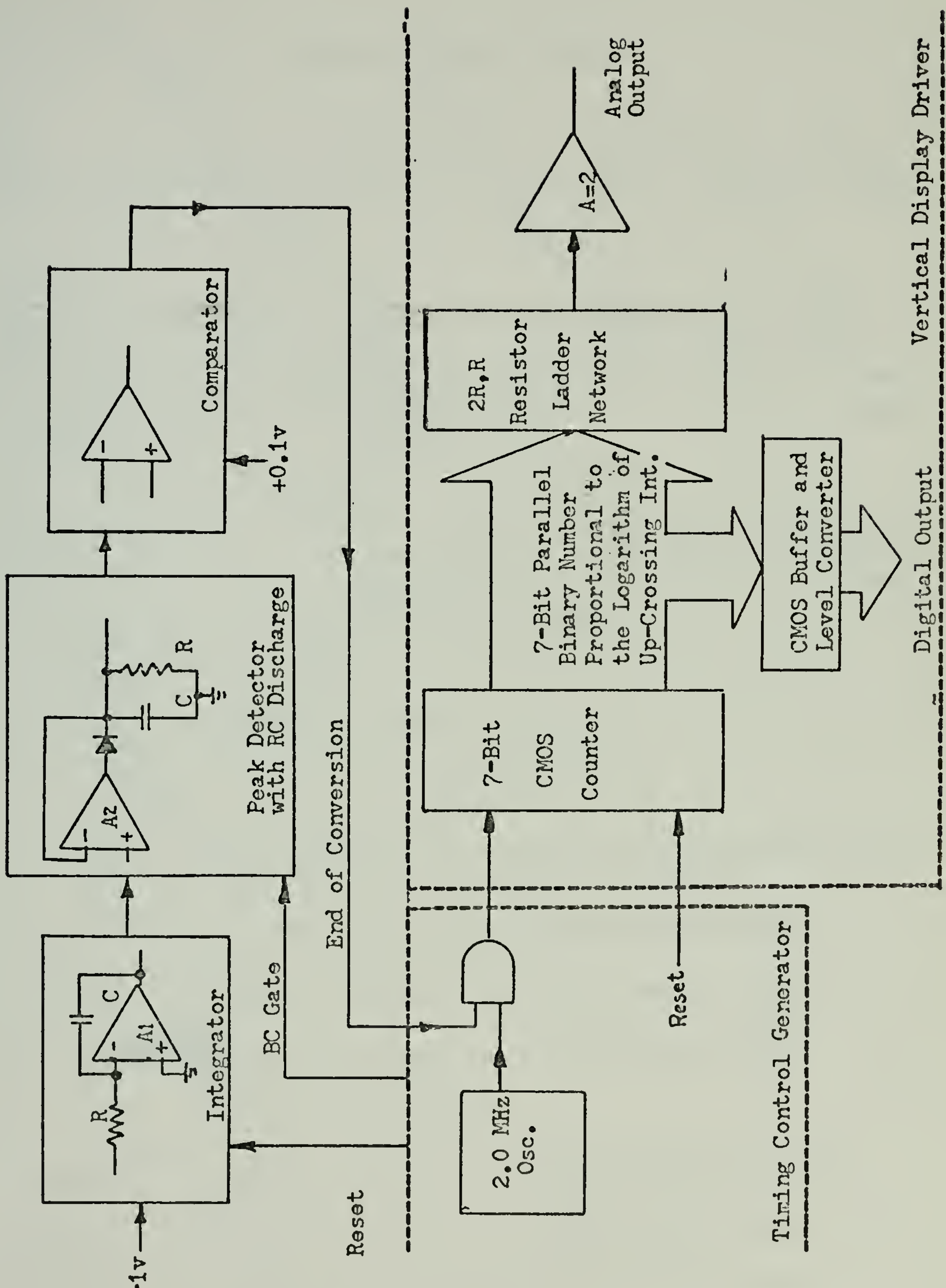


Figure 7a. Log RCI Converter

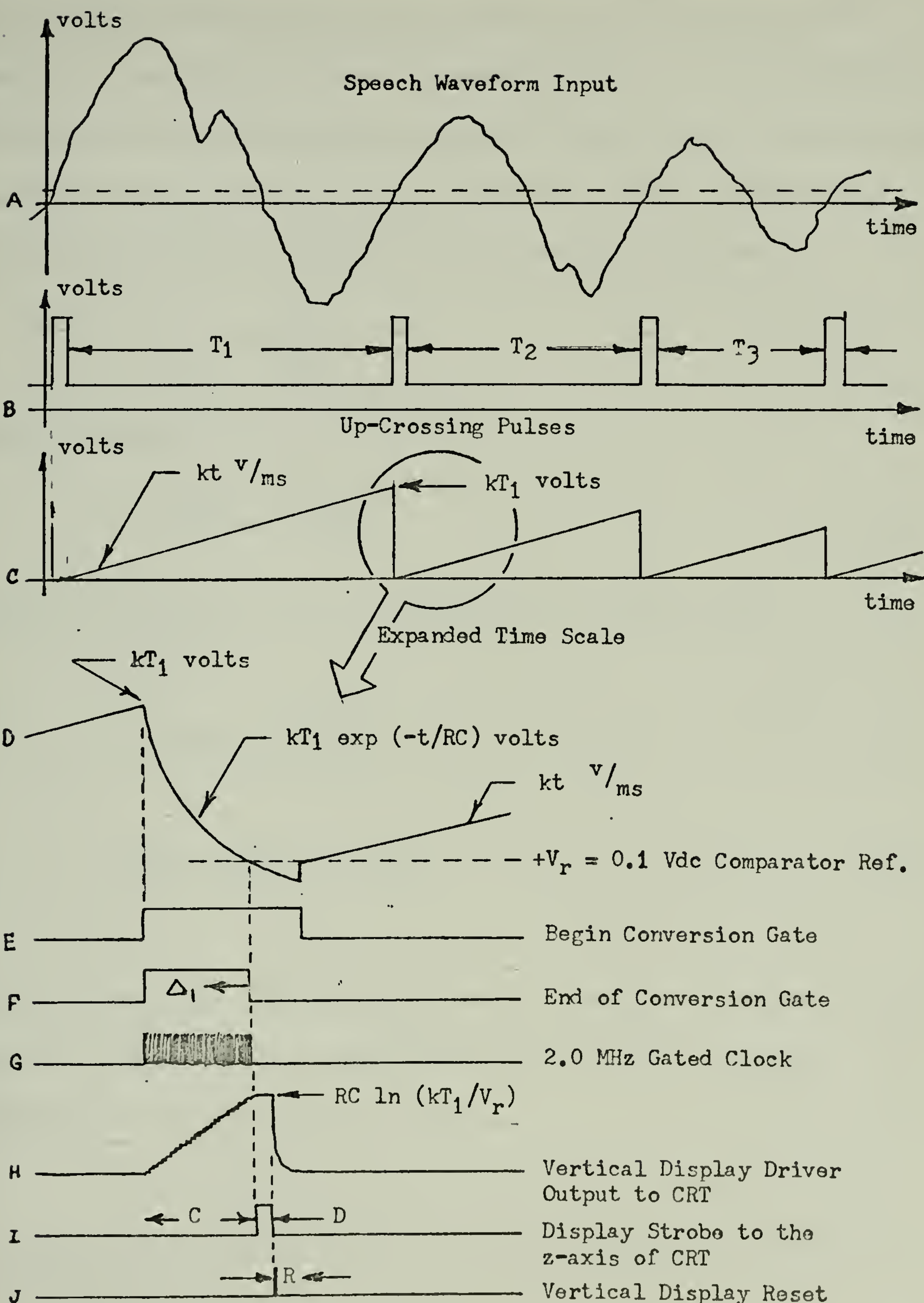


Figure 7b. Conversion Timing Diagram

capacitor is equal to the integrator output during the track mode. A 100 μsec Begin Conversion (BC) pulse, generated by the Timing Control Generator, coincident with each up-crossing pulse, sets the track-hold circuit into the hold mode for the duration of the BC pulse. During the hold mode the hold capacitor discharges through the resistor (waveform D of Figure 7b) according to the well known exponential law given by

$$v_c = v_i \exp\left(-\frac{t}{RC}\right) \tag{5}$$

Substituting v_i from (4) into (5) yields

$$v_c = kT_n \exp\left(-\frac{t}{RC}\right) \tag{6}$$

Taking the logarithm of (6) and rearranging the equation gives

$$t = RC \ln\left(\frac{kT_n}{v_c}\right) \tag{7}$$

$$RC = 13.9 \times 10^{-6} \text{ sec.} \tag{8}$$

where $v_c = V_r$ and the time constant, RC, is found by solving equation (7) with $t = t(\max) = 64$ μsec (maximum desired conversion time), $T_n = T_n(\max) = 10$ msec, and $V_r = +0.1$ Vdc.

Over the input frequency range of interest, 100 Hz to 10 KHz, the peak ramp voltage generated by integrator A_1 is defined by

$$0.1 \text{ Vdc} \leq kT_n \leq 10.0 \text{ Vdc} \tag{9}$$

where

$$0.1 \text{ msec} \leq T_n \leq 10.0 \text{ msec} \tag{10}$$

and combining equations (7), (8) and (9) yields the range of t as

$$0.0\ \mu\text{sec} \leq t \leq 64.0\ \mu\text{sec}. \tag{11}$$

Equation (7) is an analytical expression of the linear-to-logarithmic transformation of the up-crossing interval of the input acoustical waveform. The expression shows that the decaying capacitor voltage becomes equal to the reference voltage, V_r, after a time, t, that is related to the initial capacitor voltage, kT_n, by a logarithmic relation.

The reference voltage, V_r, and the exponentially decaying capacitor voltage are applied to a differential comparator (Figure 7a). The output of the comparator is, therefore, a logic "1" for a duration equal to t. This logic signal is used to gate a 2.0 MHz clock to the Vertical Axis Counter-D/A Converter in the Vertical Display Driver circuit (Figure 8). The number of clock pulses, N, gated to the counter-D/A converter in time, t, is directly proportional to t and is given by

$$N = 2t \; ; \qquad t \text{ in } \mu\text{sec}. \tag{12}$$

and from equations (7) and (12)

$$N_n = 27.8 \ln\left(\frac{kT_n}{V_r}\right) \quad ; \quad n = 1, 2, \cdots, m \tag{13}$$

where **N** is an integer formed by truncating the decimal fraction.

6. Display Driver

The 2.0 Mhz clock input from the Timing Control Generator (waveform G, Figure 7b) is applied directly to a CMOS seven-bit counter for accumulation during the gating interval. The parallel counter outputs are brought out for use in follow-on circuits. The CMOS counter selected for this application has buffered outputs which are satisfactory switches for driving a 2R, R resistor ladder D/A decoder. [28] The ladder network is connected directly to the counter digital outputs as illustrated in Figure 8. The output of the 2R, R D/A network is connected to a unity gain inverting operational amplifier which drives the vertical input of a storage oscilloscope. The analog output of the Vertical Display Driver (Figure 8) given by

$$V_D = 2R\left[\sum_{n=0}^{6} \frac{A_n}{6R(2^n)} V_{REF}\right]. \tag{14}$$

where

$$A_n = \begin{Bmatrix} 1 \text{ if nth bit is ON} \\ 0 \text{ if nth bit is OFF} \end{Bmatrix} ;$$

$$R = 150\text{ K} ;$$

$$V_{REF} = V_{dd} \text{ of the CMOS counter.}$$

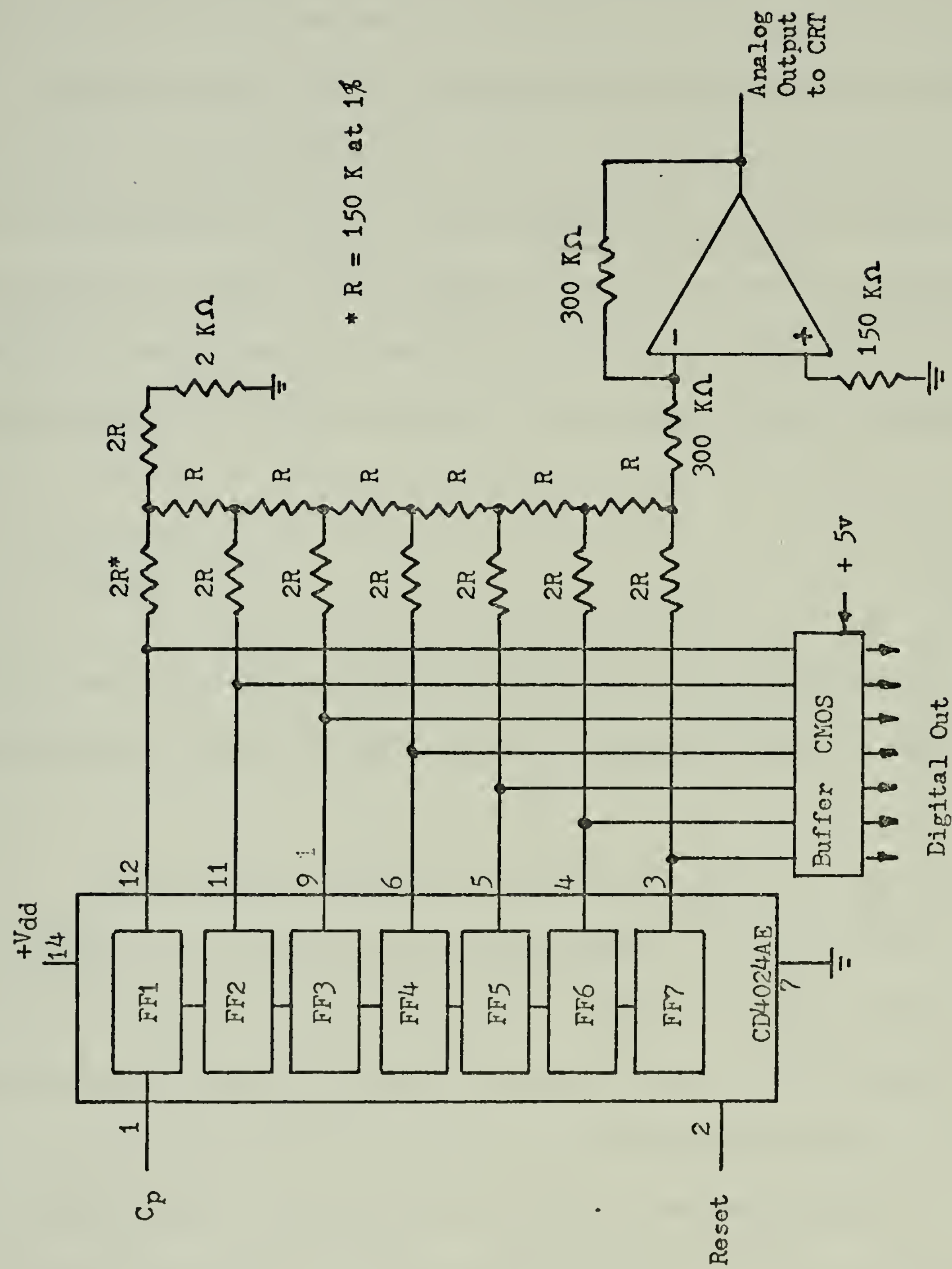


Figure 8. Vertical Display Driver

is, therefore, a voltage proportional to the Logarithm of the Reciprocal of the Up-Crossing Interval (Log RCI).

Although the vertical axis of the CRT display has the physical dimensions of volts, it is desirable to redefine the axis in dimensionless units. Since the display is similar in appearance to the well known voice spectrogram, it seemed appropriate to label the vertical axis on a logarithmic two-decade scale from 100 to 10,000 in order to further enhance this similarity. However, one should not consider the scale as a measure of frequency since it bears no demonstrable analytical relationship to the spectral content of the voice waveform. This arbitrary parameter is defined by

$$\Delta_n \propto N \qquad (15)$$

where the vertical coordinate of the parameter may be determined analytically by expressing N_n in equation (13) as a binary number and applying equation (14).

An additional justification for labeling the axis in Δ_n units is in allowing one to visualize the displayed pattern in terms of relative frequency content with the logarithmic Δ, scale increasing vertically. This is a standard display technique in spectral analysis and, therefore, enables the eye-brain system to function in an environment to which it has become accustomed. Furthermore, it facilitates describing the display as showing low, intermediate, or high frequency content in the voice waveform since there is this gross correspondence between the Δ scale value and the spectral range of the speech waveform.

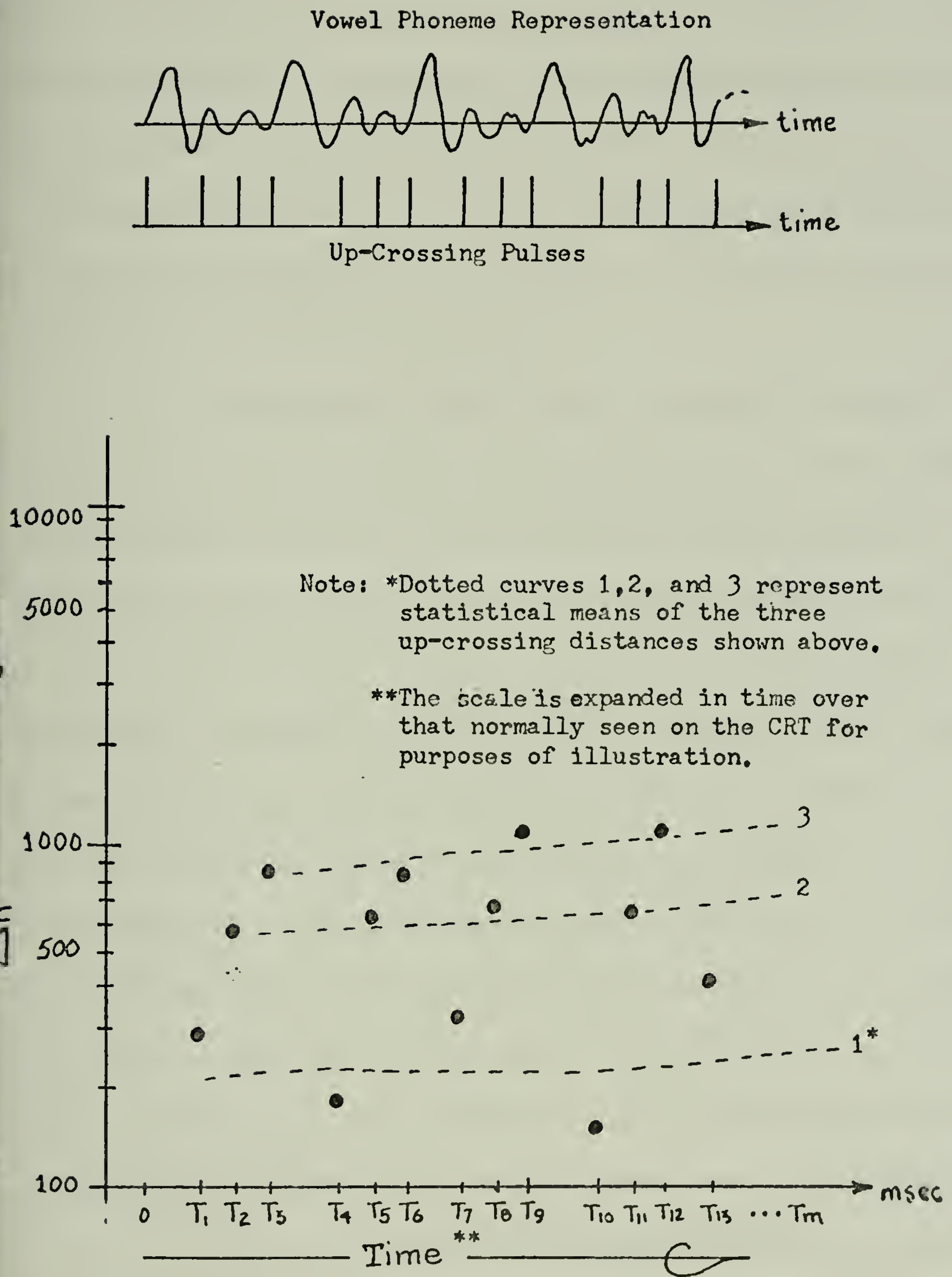


Figure 9. Representative Display Generation (vowel)

The CRT presentation displays increasing Δ along the positive vertical axis and time or accumulated up-crossings along the positive horizontal axis. The CRT vertical zero-voltage axis is set at the vertical position corresponding to 10,000 Δ (10 KΔ) which is normally the top of the CRT display and the beam is deflected vertically downward as the conversion process proceeds. At the end of the conversion, the beam is unblanked by the z-axis Display Strobe for 10 μs and a discrete dot is displayed at the appropriate Δ and the horizontal axis coordinates as illustrated in Figure 9.

The Horizontal Display Driver is identical in concept to the vertical Display Driver except it is constructed using a 12-bit CMOS output-buffered counter and drives a voltage-follower amplifier. Provisions are made to use an internal variable frequency oscillator or the up-crossing pulses as the input clock for this counter D/A converter. With the oscillator as an input, the driver generates a positive staircase voltage of up to 4096 steps within a real-time determined by the oscillator frequency. The staircase is used as an incremental real-time sweep voltage for the horizontal axis of the CRT. With up-crossing pulses as the input, the driver generates an identical staircase with accumulated up-crossings as the variable rather than time. In both cases the staircase sweep is initiated by the Timing Control Generator when the VOX signal from the AGC/Squelch Amplifier is turned ON and terminates (resets) the sweep when the VOX signal turns OFF.

The usefulness of this incremental sweep over a linear analog sweep becomes apparent from the unique display modes that may be derived from the binary output of the counter. A direct read-out of the binary number

would provide an indication of the time of occurrence within an utterance of a particular feature observed in the display while an arithematic subtraction of two such numbers would give the relative time between two features. Thumbwheel switches and digital comparator logic would allow selection of a variable length window anywhere within the time of the utterance for expansion in the horizontal axis. This would allow better visual comparison for display features and would be particularly useful for analysis of prerecorded speech.

III. RESULTS

The major thrust of this thesis has been in the development and instrumentation of the time-domain analysis technique previously discussed. The basic display system provides a low cost research tool for investigation, in the time-domain, of the acoustic properties of the consonant phonemes (stops and fricatives).

Although the data base is small, studies of these displays have revealed readily distinguishable visual patterns, which are useful for discriminating some consonants, often even in connected speech.

The reason for not averaging up-crossing is that in the speech waveform itself there are significant acoustic features which are only one or a few cycles in duration. If cycles are averaged these features are irrevocably lost. Such transient events frequently occur at vowel-consonant and consonant-vowel boundaries [21] as well as between other acoustically distinct regions, within stop consonants for example.

For example, one such feature often occurs at the transition from a stop or fricative to a following vowel. There appears to be a relatively long and intense cycle between the consonant and vowel. Sometimes there are several cycles before the vowel. On the display this feature appears as relatively "low frequency" dot or group of dots immediately preceding the vowel. The occurrence of these transition cycles coincides with the up-swing in energy from the consonant to the vowel.

Another area where relatively consistent time-domain features have been observed is during the course of the stop constants. The stop constant consists of three distinct regions; the initial pause, a release, and aspiration. The pause is characterized in the speech waveform as a region of very low energy, irregular acoustic activity which is terminated abruptly by the release characterized by many greater amplitude, high frequency cycles. In the up-crossing displays, the initial pause appears as either one or a few outstanding "low frequency" dots immediately preceding the release activity.

Due to monetary limitations extensive hard-copy visual displays have not been produced. However, representative samples have been taken to illustrate the basic utility and potential of the display for speech research in the quest for the acoustical invariants of human speech.

It has been found that people with no prior experience in speech research could often correctly locate and identify certain consonant phonemes in unlabeled displays with the information given above. No formal or extensive experiment was performed due to limited hard-copy displays.

It should be pointed out that the ability of a human observer to distinguish these phoneme patterns in no way ensures that an automatic system can be developed to do as well. However, several applications of the basic display system which involves a human observer in a feedback link are discussed in the conclusion section of this thesis.

Representative samples of the data taken with a Polaroid scope camera from the CRT of a Tektronix Model 141 storage oscilloscope are presented in Figures 10 - 17. All displays were generated from tape recorded

utterances recorded on an Akai model 300 professional magnetic tape recorder (frequency response 50Hz - 15Khz at ± 3 db).

A dot display of the phoneme /s/ (unvoiced fricative) spoken in isolation (no preceding or trailing phoneme) is shown in Figures 10 and 11. Figure 10 represents the extremes of speaker variations (among five male speakers) as determined by visual inspection of the displays. Figure 11 represents a composite of all five speakers produced by electronically superimposing the displays. The VOX signal for each utterance was used to gate clock pulses to the Horizontal Display Counter Driver without erasing the utterance previously stored on the CRT. The optical integration serves to enhance gross similarities in the visual patterns. Particular attention is called to the band-structure of the displays which are produced by the cavity shaping of the acoustic noise. This characteristic is particularly insensitive to speaker variation (male-female).

The phoneme /er/ (voiced fricative) is shown in Figures 12 and 13. A composite display could not be produced from these recorded utterances due to microphone noise immediately preceding each utterance. However, these displays are presented to show the display evolution from unvoiced fricative to voiced fricative to semi-vowel. This phoneme exhibits consonant characteristic in the initial portion of the utterance and vowel-like periodic structure in the trailing steady-state region. The transition is very apparent in the time-domain display. (The display time axis is 10 msec/cm.)

The semi-vowel /ℓ/ is shown in Figures 14 and 15. Figure 14 represents the variational extremes among five male speakers while Figure 15 is the composite integrated display of all five utterances. The /ℓ/

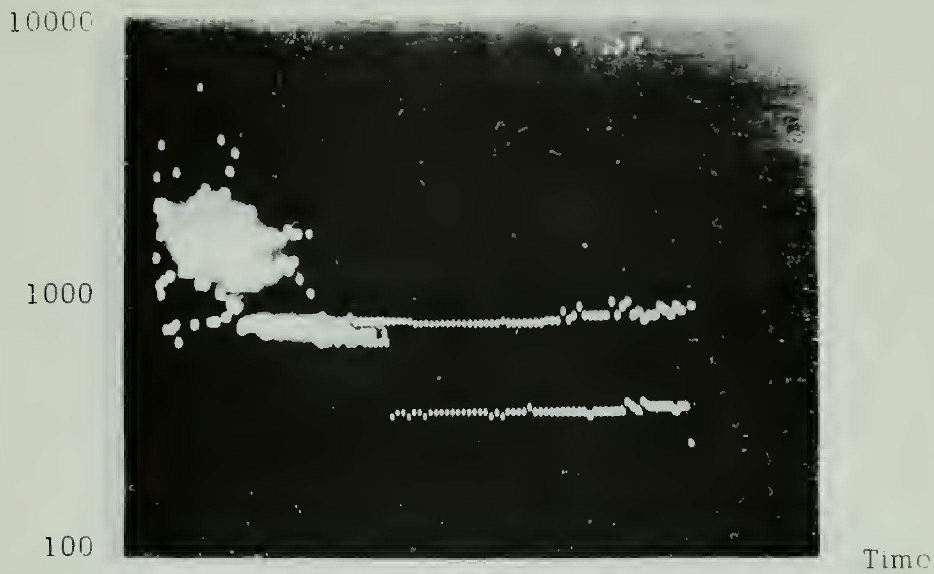
Phoneme /*s*/ by speaker HR

Phoneme /*s*/ by speaker LE

Figure 10

Composite of five male speakers
for the phoneme /s/

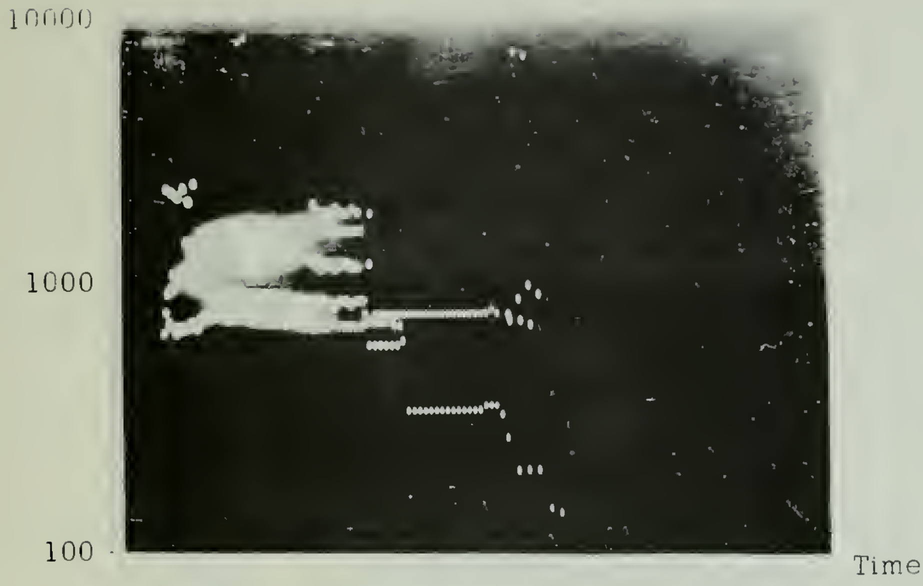
Figure 11



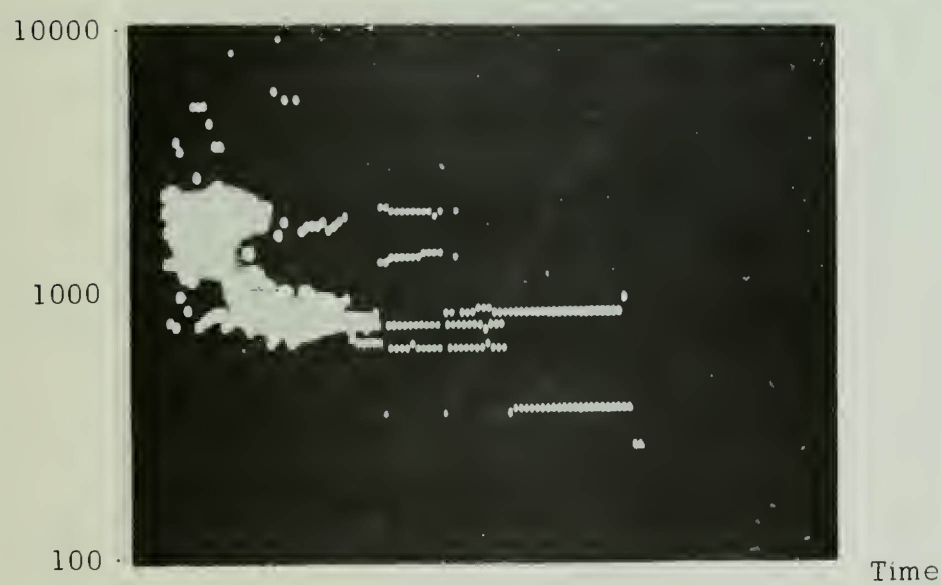
Phoneme /**er**/ by speaker MK

Phoneme /**er**/ by speaker LE

Figure 12

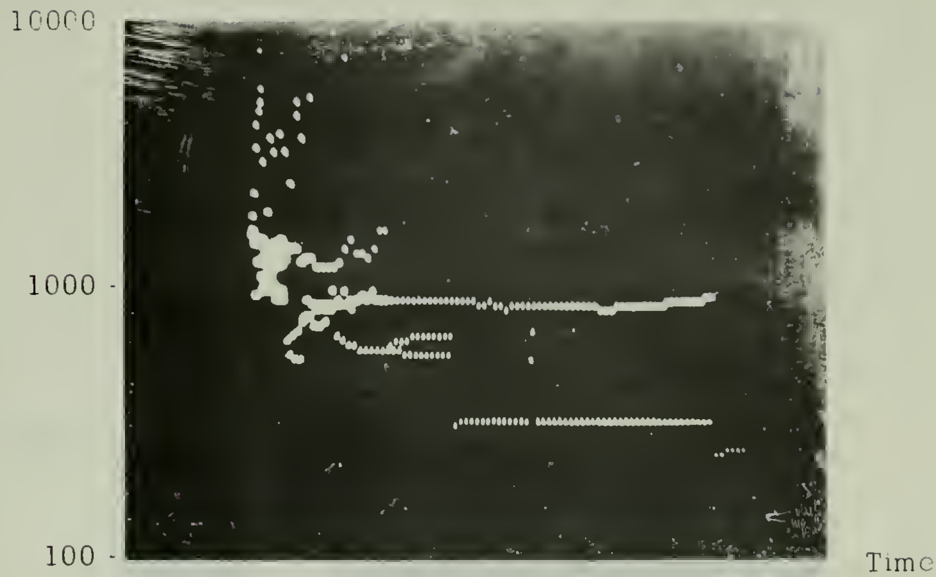


Phoneme /*er*/ by speaker GD

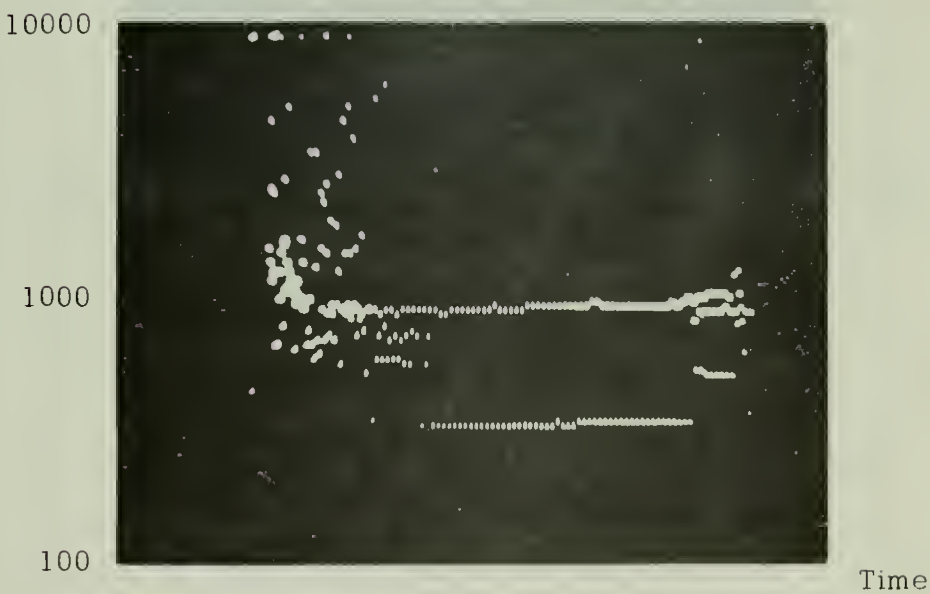


Phoneme /*er*/ by speaker DM

Figure 13



Phoneme /ℓ/ by speaker MK



Phoneme /ℓ/ by speaker LE

Figure 14

Composite of five male speakers
for the phoneme /ℓ/

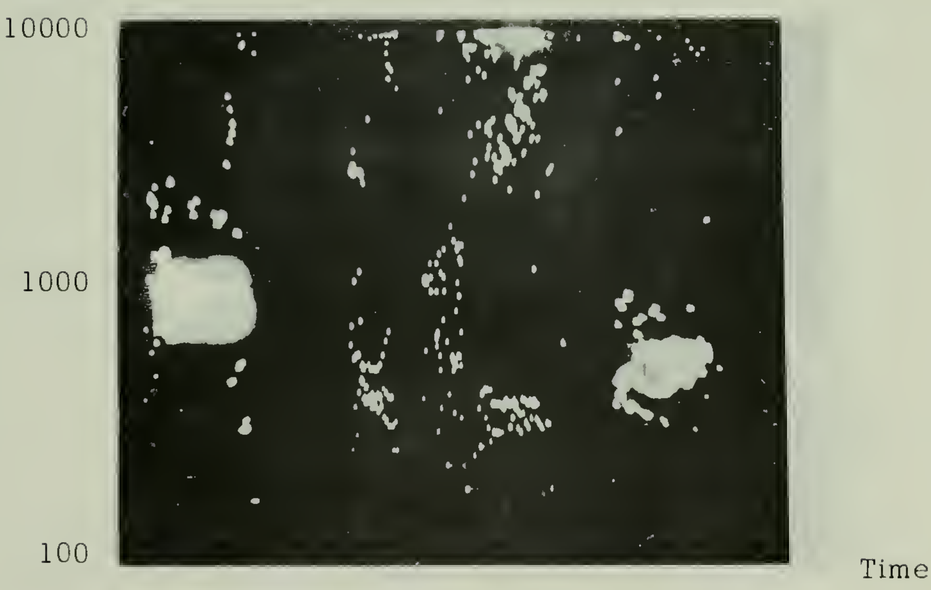
Figure 15

phoneme shows essentially vowel-like structure with some initial consonant activity. All voiced displays are pitch sensitive in that the constant Δ "bars" shift with changes in pitch-frequency. This is consistent with the formation of vowel acoustic waveforms as discussed in the recommendation section and illustrated by Figure 19.

Figure 16 shows the display pattern of the utterance "Pawn to queen four" spoken by two male speakers. The gross pattern similarities are readily apparent as are the natural breaks before stop consonants /P/, /t/, and /K/ (in queen).

Figure 17 is the display of the phrase "Oh my aching back" spoken by two male speakers. These displays have been "normalized" by making the total horizontal display length equal for all phrases. The threshold adjustment in the Logarmithic Threshold Crossing Detector has been set to eliminate the low level signals of the fricative sounds (center clipping). This illustrates the effects of the threshold adjustment on display presentation. The words in this phrase were intentionally stressed to produce well defined vowel patterns as compared to the phrase in Figure 16 where the phrase was spoken normally. Note that the stop consonant /K/ (marked by white arrows) in both displays shows an initial pause and few "low-frequency" dots followed by a cluster of "high-frequency" dots. While the time relation is difficult to see on this phrase display, expanded displays have shown the time sequence readily.

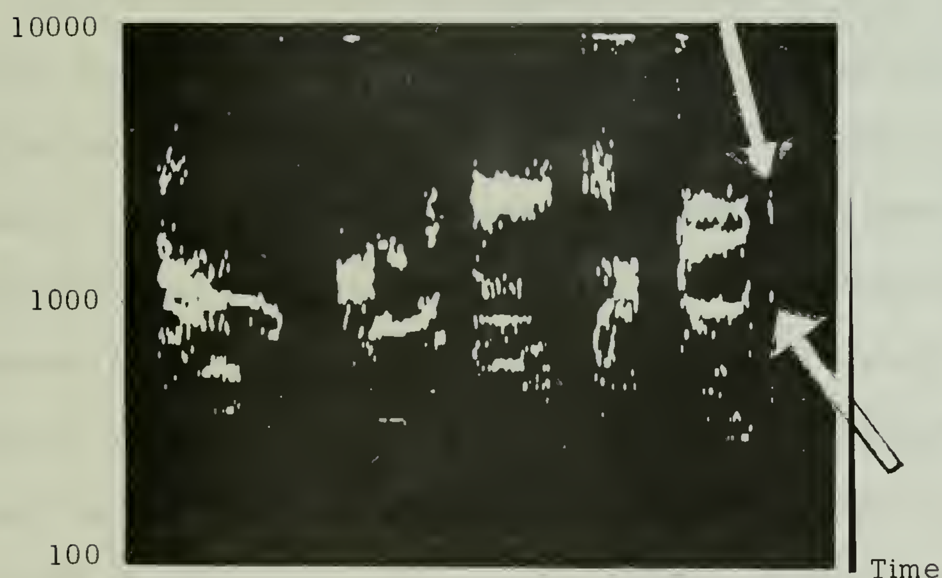
"Pawn to Queen Four" (GD)



"Pawn to Queen Four" (HR)

Figure 16

"Oh My Aching Back" (MK)



"Oh My Aching Back" (DM)

Figure 17

IV. RECOMMENDATIONS FOR FUTURE RESEARCH

A. A PROGRAM FOR COMPARING SPEECH RECOGNITION TECHNIQUES

There are many factors relating to the evaluation of speech as a communication media between the severely handicapped individual and the machines and devices that influence his daily life-style. Among these factors, one of the most difficult and pressing problems concerns the development of adequate automatic speech recognition equipment. The value of a particular voice input system will depend heavily upon the success of the speech recognizer. Obviously, all the advantages of voice input systems may be overshadowed by any inadequacies in speech recognition devices.

Speech training for the handicapped is a long and highly repetitious process requiring an expensive one-to-one student-teacher relationship. The teacher often being reduced to playing the role of a speech recognizing automaton. Speech recognition systems, in contrast, are specifically designed to remove the intelligent human from the perceptive link, and to substitute for him a speech recognizing automaton. The success of the automatic speech recognizer in this application is thus established by how closely the recognizer duplicates the behavior of the sensory and perceptual processes in the human ear-brain system. (This should not be taken to mean that the recognizer must duplicate the exact mechanisms, or techniques by which the human understands speech. It does mean, however, that the

recognizer must identify an acoustic feature as the same feature perceived by the ear-brain system.)

In the following discussion only those programs concerned with speech recognition processes in the time-domain will be considered.

1. Selecting Speech Time-Domain Parameters

Speech recognition may reasonable be divided into the separate tasks of first extracting significant parameters from the speech signal and then recognizing or classifying the parameter patterns. There are many acceptable techniques for the second task of pattern classification, which will not be discussed here [1, 29, 20] . Speech parameter extraction has repeatedly been recognized as one of the major recognition problems to be overcome. If a good selection of parameters or features is extracted, then, in most cases, any of several recognition algorithms may be used to achieve correct recognition. On the other hand, no amount of recognition elegance can make up for an inadequate set of measured parameters.

If one is free to select as many measurements or parameters as he wishes, somewhat careless selection of the parameters may be possible while still preserving in the parameter set those which are most important or "information-carrying". However, economy usually dictates that the set of measured parameters be as small as possible, so that one must carefully select parameters to be measured. As Nilsson [31] has observed:

> "Unfortunately, there is very little theory to guide out selection of measurements. At worst this selection process is guided solely by the designer's intuitive ideas about which measurements play an important role in the

classification to be performed. At best the process can make use of known information about some measurements that are certain to be important".

In speech recognition, (and in the absence of a theory to guide parameter selection) the most common extraction technique has been to obtain the short-term frequency spectrum of the speech signal and to pick out important spectral highlights such as formant frequencies and their time changes through phonemes, syllables, and longer utterances. The formants are closely related to resonances of the vocal tract cavities; and pitch, voicing and other articulation phenomena are evident in the spectrum. These factors justify, within the speech production viewpoint, some of the long history of spectrum and spectrogram approaches to speech analysis.

However, as discussed before, thirty years of research and practice with speech spectra and spectrograms [32] have not led man much closer to achieving the goal of automatic speech recognition. Several reasons for this inadequacy of spectral approaches could be given. From a sensory viewpoint, one can argue that the ear does not, in anything more than a very loose sense, perform a spectral analysis [33]. From a perceptual viewpoint it can be argued that the important linguistic or perceptual information is effectively buried in the large amount of information available in the spectrum. One can also observe that the short-term spectrum is only an approximation to the Fourier series, which in turn is only an approximation to the proper Fourier integral; the filters traditionally used in spectrum analysis add their own information to the speech spectrum; and much of the important time and transient information is lost in a transformation to spectrum form.

These and other factors have suggested the need to closely examine alternatives to the spectrum approach for speech parameter extraction. Alternative extraction techniques, which vary considerably in the extent to which they deviate from standard spectral analysis, have been proposed. Some of these alternative extractors attempt to provide a minimal set of simple measurements which are most significant to automatic speech recognition; thereby, significantly simplifying the extractor equipment and recognition algorithms.

2. Comparatively Evaluating Parameter Extractors

With the development of a variety of speech parameter extraction techniques, an important problem has arisen. How does one compare the various extractors in order to judge their relative effectiveness for speech recognition and to understand the distinct situations in which one extractor serves better than another?

One method for comparatively evaluating time-domain parameter extractors and relating their results to the design of better systems is suggested by Figure 18. Each extractor shown in Figure 18 may be implemented with minimal hardware by utilizing the digital and analog signals available from the time-domain display/analyzer described in this thesis. The central processor for this system is not a computer but the human eye-brain system -- a proven pattern recognizer. It must be pointed out that the ability of the human observer to recognize patterns in no way ensures that a machine can be constructed to do likewise. However, this ability can prove invaluable to basic research in speech time-domain analysis.

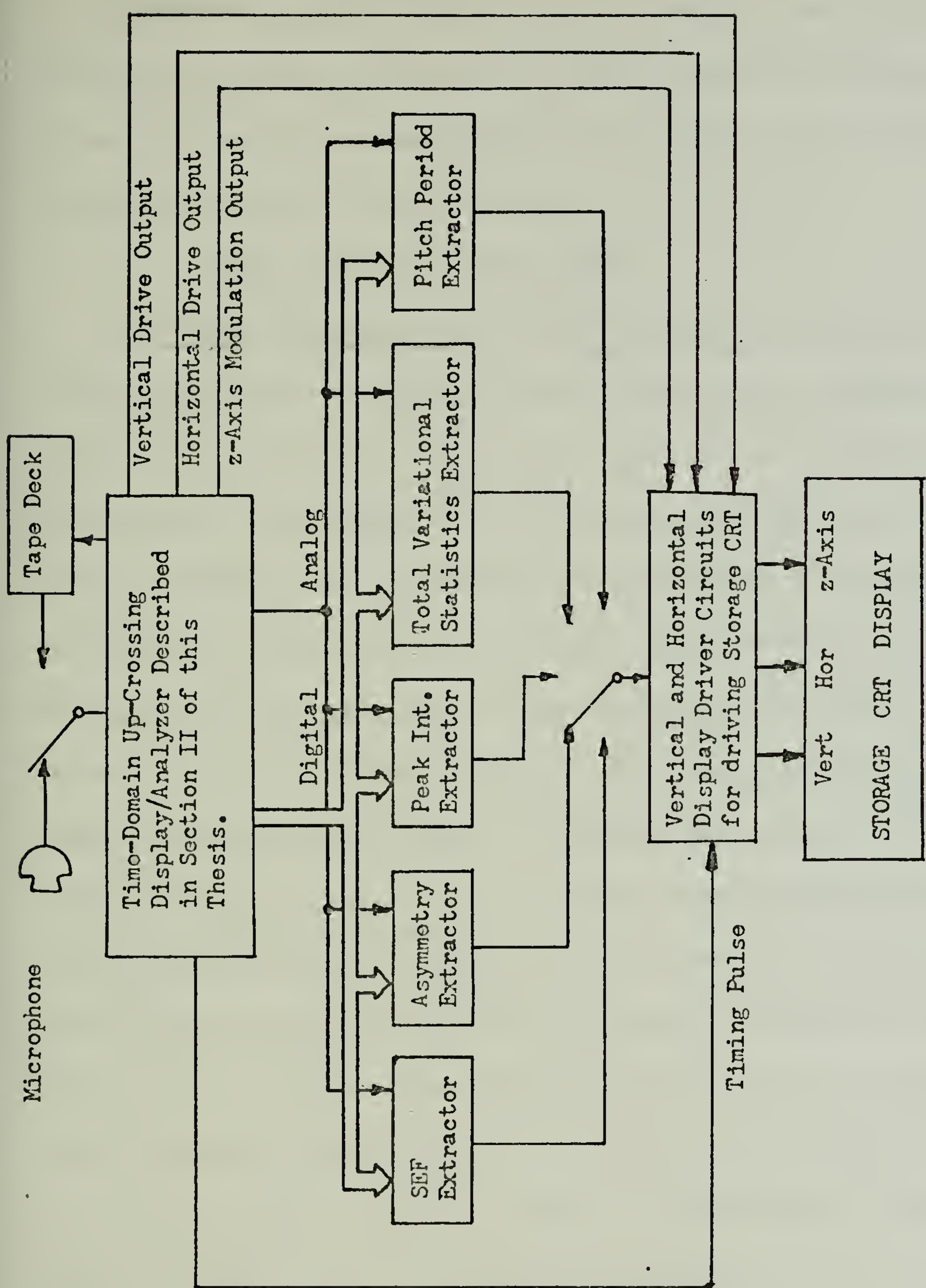


Figure 18. Block Diagram for a System to Comparatively Evaluate Parameter Extractors.

B. SEVERAL EXAMPLE EXTRACTORS

A complete exposition of the various extractors shown in Figure 18 will not be attempted in this thesis. Instead a conceptual summary of the technique will be presented along with the basic methods of obtaining the parameters from the display/analyzer.

1. Single Equivalent Formant (SEF)

The extraction of the SEF is based upon characteristics of the speech waveform and the fact that when a person hears a multiformant sound, his attention focuses only on the dominant one. The presence of other formant (called recessives) only serve to modify the perceived sound slightly from that of the dominant formant. The formation of a three formant speech sound is shown in Figure 19. Each gottal or noise excitation produced by the vocal organs acts as a driving function that initiates a complex ringing phenomenon in the oral cavaties. The ringing frequencies of these damped sinusoids are the formant frequencies whose energies add to produce the complex speech wave. The relative formant amplitudes determine how these waveforms add. Figures 20 (a), (b), and (c) show this in detail. The third formant (highest frequency) can usually be neglected in comparison to the first two. When the first formant is larger in amplitude than the second, the period of the first major oscillation of the complex speech wave is approximately equal to the period of the first formant. Similarly when the second formant is dominant in amplitude the period is approximately the same as the second formant. When the two formants are approximately equal in amplitude, the resultant period is equal to the average of the two formant periods.

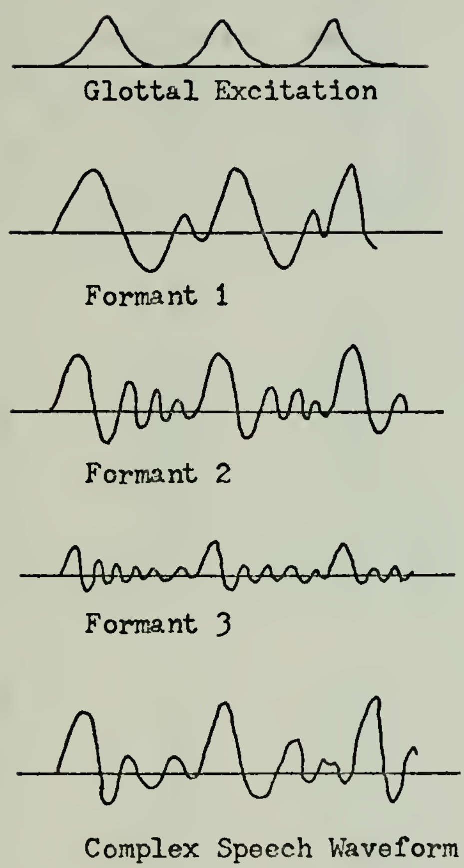


Figure 19. The Formation of Three-Formant Speech Sounds.

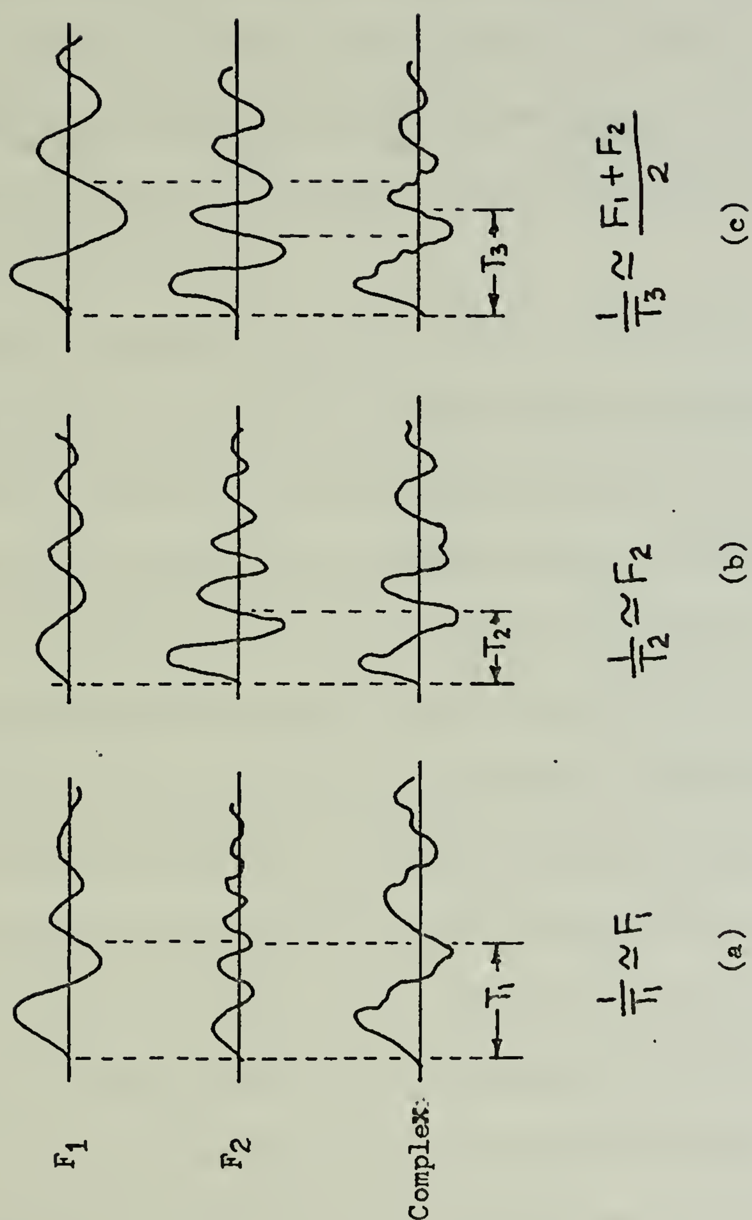


Figure 20. Single Equivilent Formant Frequency determination from Complex Wave.

Focht [34, 35] has shown that for the back vowels, such as U in "tool" and ɔ in "talk", the first formant is dominant and nearly equal in frequency and amplitude to the SEF. For the front vowels, such as ɛ in "ten" and I in "tip", it is the second formant that is dominant and nearly equal in amplitude and frequency to the SEF. The first two formants for the central vowels, such as ɑ in "tar," are nearly equal in amplitude and the SEF frequency is approximately the average of the first and second formant frequencies. The third formant has little or no effect on the SEF (Figure 21).

Focht also found that the period of the first oscillation of the complex speech waveform after each glottal excitation provides a measure of the SEF period. Therefore, to extract the SEF frequency from the complex speech waveform, the period of this oscillation must be measured.

From the Theory of Operation of the Display/Analyzer, it is noted that each period of the complex speech waveform is successively measured and parameterized into a Log RCI parameter. This parameter is available as an analog voltage or as a seven-bit binary number. Therefore, a parameter which is the logarithm of the SEF frequency may be extracted if the time of glottal excitation is known. The pitch pulse output of the Pitch Extractor (described below) is time-coherent with glottal excitation and thus may be used for this purpose. The Log-SEF frequency extractor then becomes a simple analog or digital switch activated by the pitch pulse. Alternatively, the display might be blanked for all up-crossing information except the Log-SEF frequency.

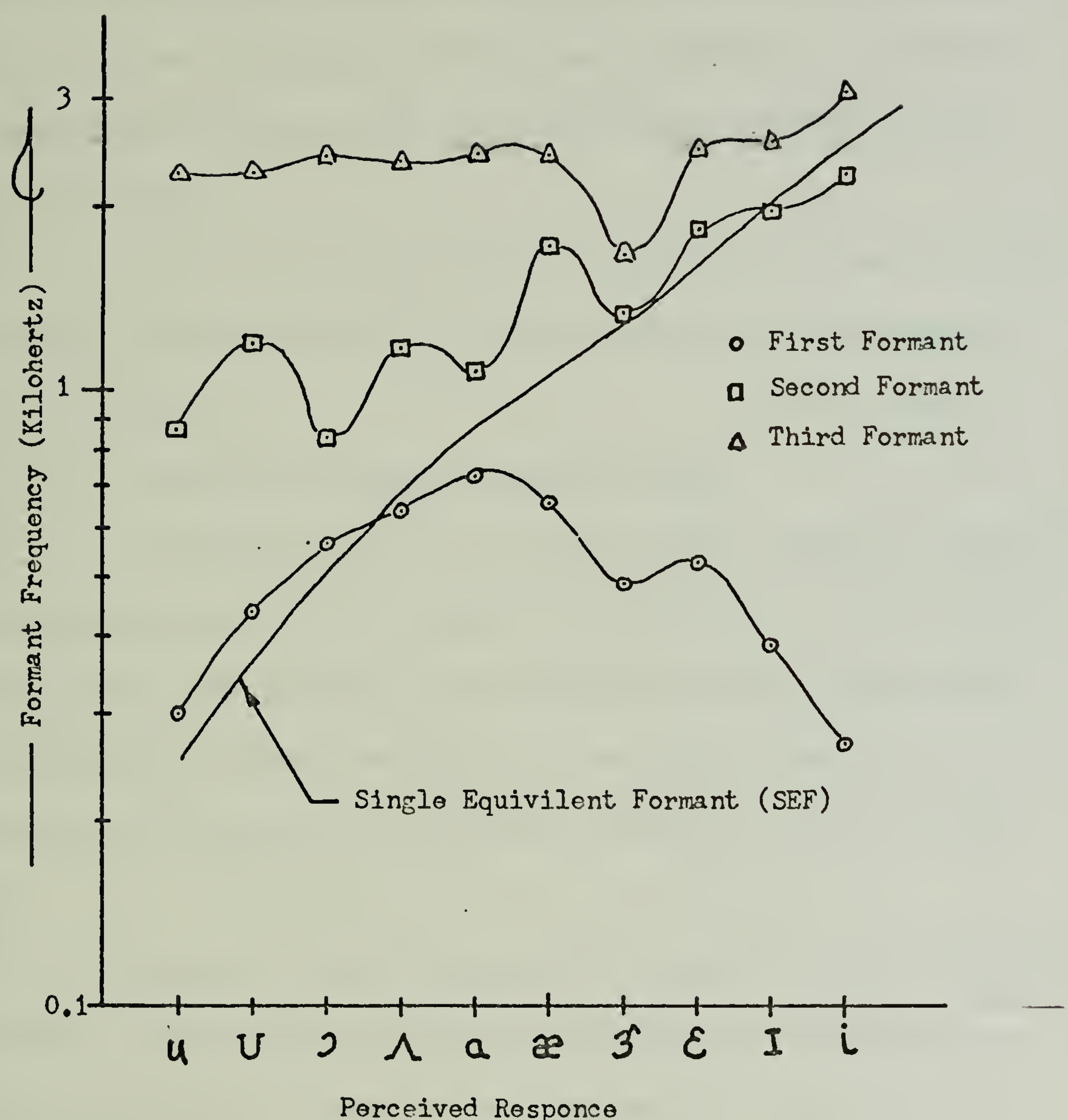


Figure 21. Correlation of the Single Equivilent Formant and Human vowel formant locations [34].

Another parameter mentioned by Focht was SEF amplitude which is the peak amplitude of the speech waveform in the first up-crossing interval after glottal excitation. This measure may be obtained from the Peak-Intensity Extractor (described in "3" below) and the pitch pulse. If so desired, this measure might be displayed as a z-axis intensity modulation concurrent with the Log-SEF frequency parameter by controlling the amplitude of the Display Strobe proportional to the Peak-Intensity Extractor output.

The display thus produced allows rapid evaluation of extractor performance and facilitates the design of sequential logic circuits for feature recognition.

2. Waveform Asymmetry of Voiced Speech

Waveform asymmetry is a measure of the difference in magnitude between positive and negative peaks of a waveform. All unvoiced speech, composed of nonharmonically related components, will be symmetrical about the base line when averaged over relatively long periods of time. Examples of phonemes with symmetrical waveforms are /f/, /th/, and /s/.

Figure 22 shows the approximate waveforms for three different waveforms. The /a/ sound shown possesses positive asymmetry since the positive peaks are larger than each succeeding negative peak. The /e/ sound has negative asymmetry. All voiced sounds exhibit asymmetry as spoken or can be modified to exhibit asymmetry by phase shifting its harmonic components. This measurement can quite accurately discriminate between voiced and unvoiced speech and, therefore, can perform the function

of segmenting a word into voiced or unvoiced portions. Segmentation of a word is very important in a limited vocabulary device since the sequence of unvoice and voiced portions aids identification.

The origin of asymmetry may be understood by considering a repetitive waveform made up of a fundamental and higher order odd harmonics. With only odd harmonics present the waveform will exhibit half-wave symmetry which is independent of the phase and amplitude relationships of the various harmonics. It is obvious that a waveform containing only odd harmonics can possess no peak asymmetry. Now consider a waveform composed of a fundamental and a second harmonic component as shown in Figure 23a. If there is no phase difference between the components, the sum of the components will be a symmetrical waveform as shown. However, if the second harmonic is shifted in phase by 45° with respect to the fundamental the waveform exhibits maximum asymmetry as shown in Figure 23b. The asymmetry is a function of two quantities: the relative magnitude of the second harmonic and the relative phase difference between the two components.

If many harmonics are present in the input waveform, the asymmetry is a very complex function of the magnitude and phases of all even harmonics relative to all odd harmonics. For most vowel sounds the amplitudes of the higher order harmonics are much smaller than the amplitude of the lower frequency components. Therefore, the waveform asymmetry is determined mainly by the fundamental and the first few harmonics. The first four components are often the major contributors to asymmetry and thus higher order harmonics may be neglected.

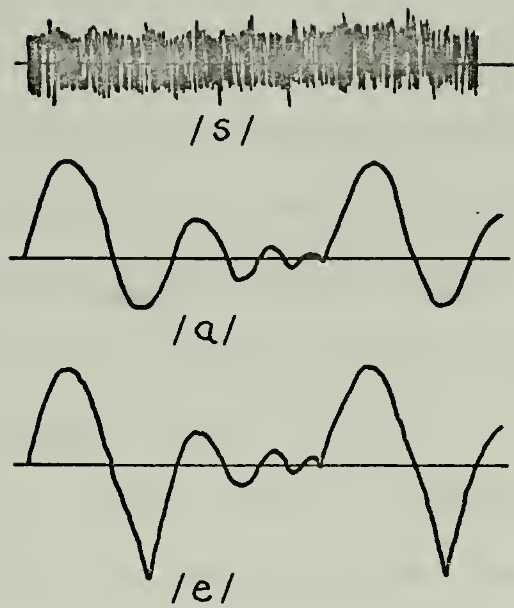


Figure 22. Waveforms of three phonemes.

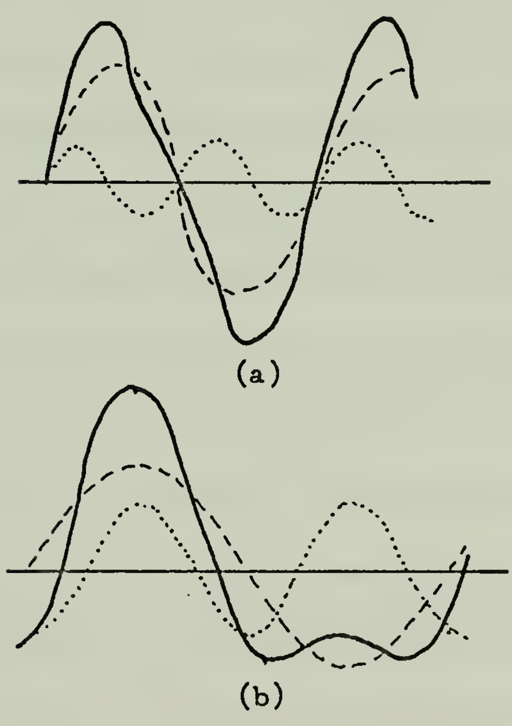


Figure 23. Origin of Asymmetry

It is reasonable to expect that the relative phases of the harmonic components will be different for different vowel sounds. The cavity of the mouth and throat is shaped differently for each vowel sound. If the exciting frequency (determined by the vocal cords) remains relatively constant as the resonant cavity changes shape, the difference between the resonant frequency and the excitation frequency changes. Since the phase shift of the cavity is strongly dependent on the difference in excitation and resonant frequencies, the components of each vowel sound will be subjected to different amounts of phase shift. Furthermore, the relative magnitudes of the components vary as the cavity is changed. Thus most vowel sounds should exhibit a different amount of waveform asymmetry. Experimental work has verified this conclusion [34] .

A simple addition to the Display/Analyzer is proposed which produces a real-time on-going display of waveform asymmetry. The operation of the circuit may best be described by reference to the block diagram of Figure 24. The positive and negative peak detector outputs from the Peak Intensity Extractor are algebracially summed and used to control the gain of a voltage controlled amplifier (VCA). The VCA controls the amplitude of the triangle wave applied to the analog gates in proportion to the difference in the positive and negative peaks of the input acoustic waveform on a cycle by cycle basis. The polarity signal from the Peak Intensity Extractor determined which analog gate is active. The output of the analog gates is summed with the respective horizontal or vertical position voltage at the input of the horizontal or vertical Display Driver Amplifier. The CRT drive signal therefore consists of the DC position voltage and a superimposed

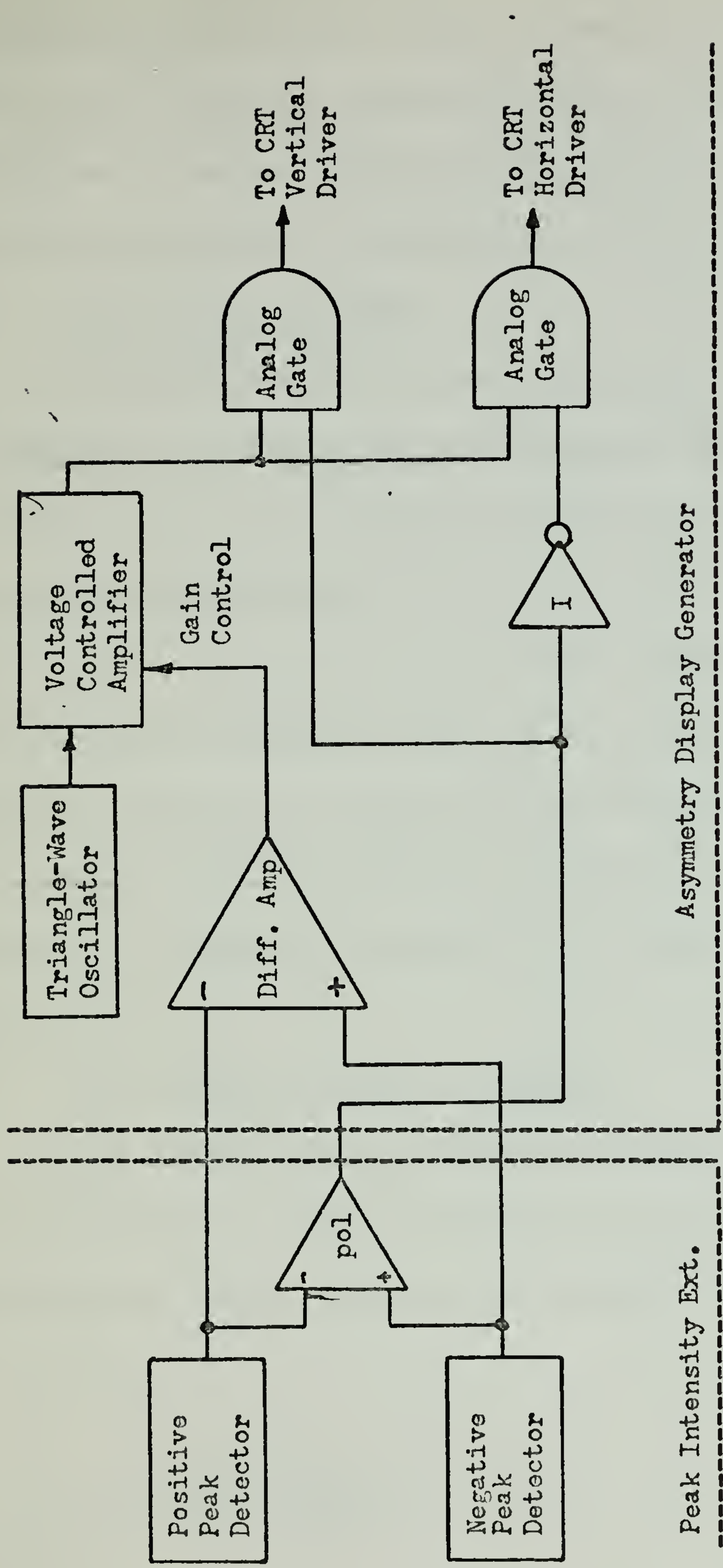


Figure 24. Asymmetry Extractor Circuit

low-amplitude triangle waveform. This drive signal will produce a display consisting of short vertical or horizontal bars whose length is proportional to the asymmetry of the input speech waveform and whose spatial position is identical to the original dot pattern display previously described.

3. Peak Intensity Extractor

A possible circuit for extracting the positive and negative peaks of the input acoustic waveforms on a cycle by cycle basis is shown in block diagram in Figure 25. The curcuit provides both analog and digital outputs for interfacing with the other extractors and with the analog display. The cyclic peak intensity of the positive and negative peaks is compared in an analog comparator and a logic level signal is developed to indicate the polarity of the signal with the greatest relative magnitude. The peak intensity might be displayed as an intensity (z-axis) modulation on the CRT or as a vertical or horizontal bar in a manner similar to the asymmetry display.

4. Total Variational Statistics Extractor

The circuit of Figure 26 has been developed as a signal-variation accumulator. The following equation states the performance of the circuit exactly, but the operations it denotes are largely implicit in the actual circuit.

$$e_o = \frac{2C_1}{C_2}\int_0^t \left|\frac{de_i}{dt}\right| dt \qquad (16)$$

After the reset switch has been momentarily closed the input-output relationship is as shown in Figure 27, the output momotonically

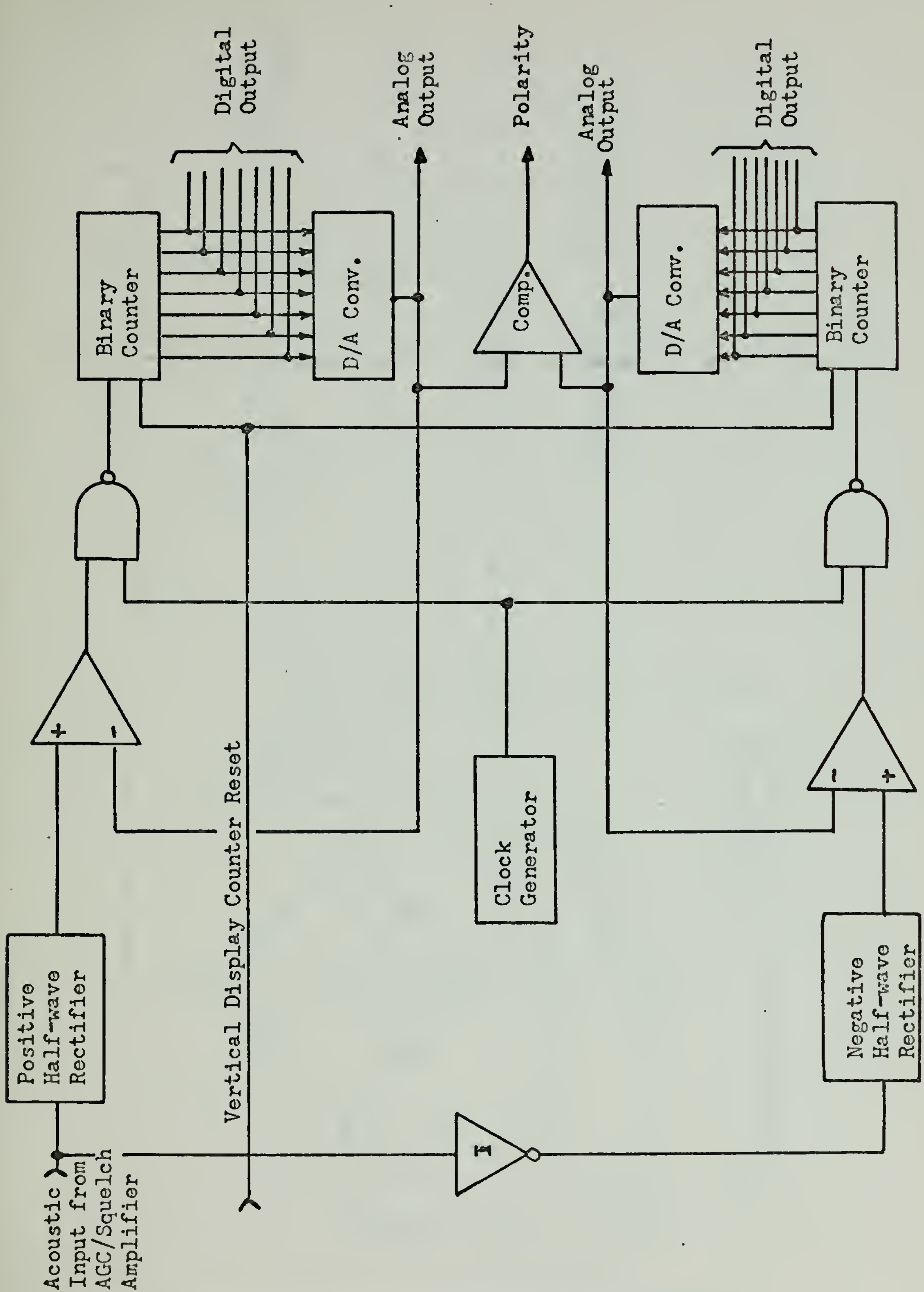


Figure 25. Peak-Intensity Extractor Schematic

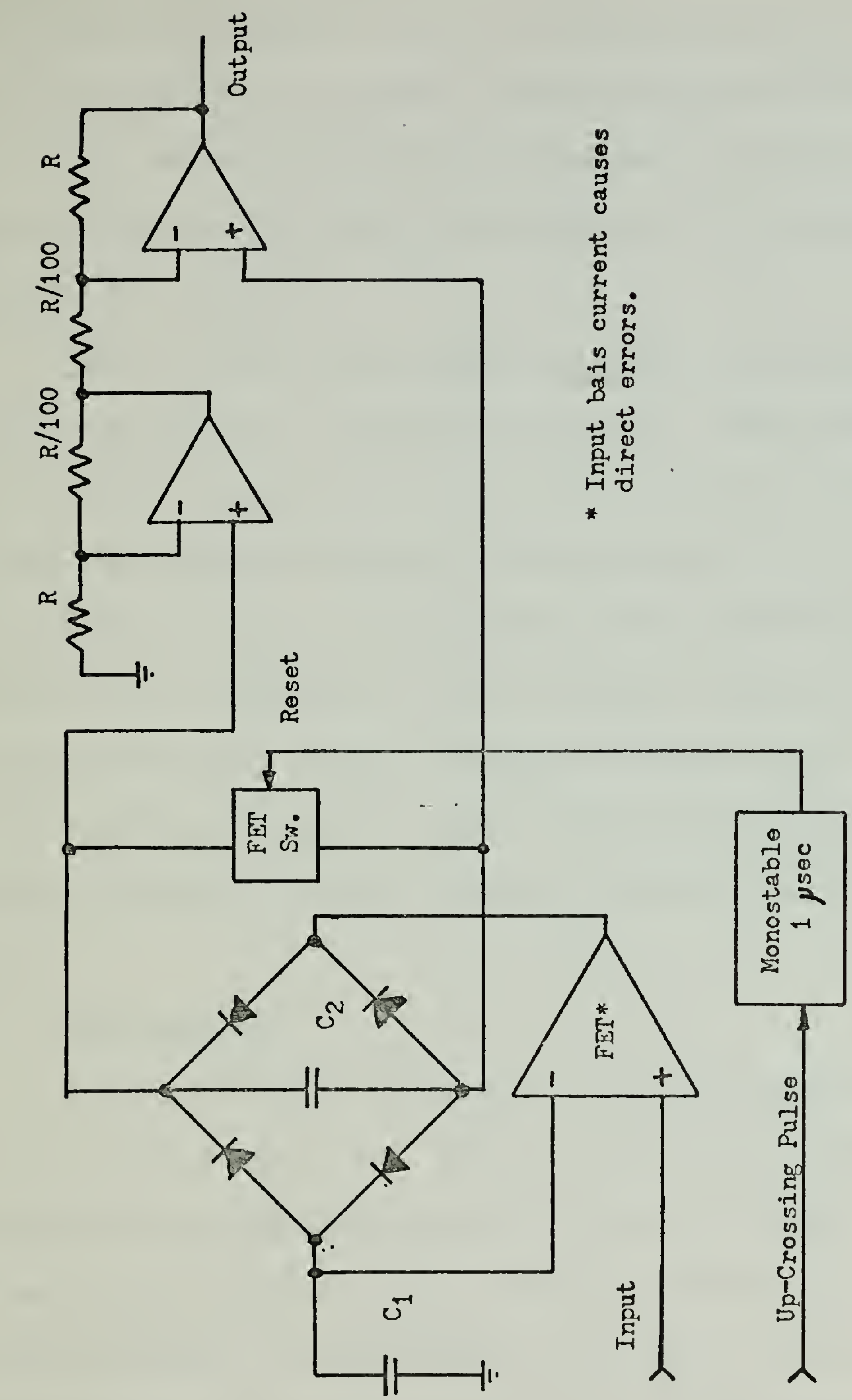


Figure 26. Signal Variational Accumulator Circuit

increasing in response to increments and decrements of the input. The output is a measure of the signal "activity" following the last reset. The circuit is reset after each up-crossing. The diode bridge causes the currents for charging and discharging C_1 in response to input signal variations to comulatively charge C_2; the voltage across C_2 is read out by the differential follower.

This circuit has shown promise as a means of segmenting the acoustic speech waveform. Voiced phonemes produce a significantly lower output voltage on a cyclic basis than does an unvoiced sound due to the broad-spectrum noise produced in unvoiced phonemes. Only very limited investigation has been conducted with less than satisfactory components used in implementing the circuit, therefore, further investigation will be required before the degree of segmentation attainable may be determined. Presently the circuit gives reliable segmentation into voiced and unvoiced phonemes, although the measure is somewhat speaker sensitive.

5. Pitch Extractor

Literature on speech analysis contains many references on techniques for pitch extraction [35 - 40] . A review of these techniques is beyond the scope of this thesis; however, one circuit which has not been covered in the literature is shown in Figure 28. This pitch extractor automatically maintains an optimum threshold with respect to an incoming acoustic signal over a 100:1 signal amplitude range. The threshold is stable with respect to the level of the acoustic signal, enabling pitch pulses to be extracted by amplitude discrimination.

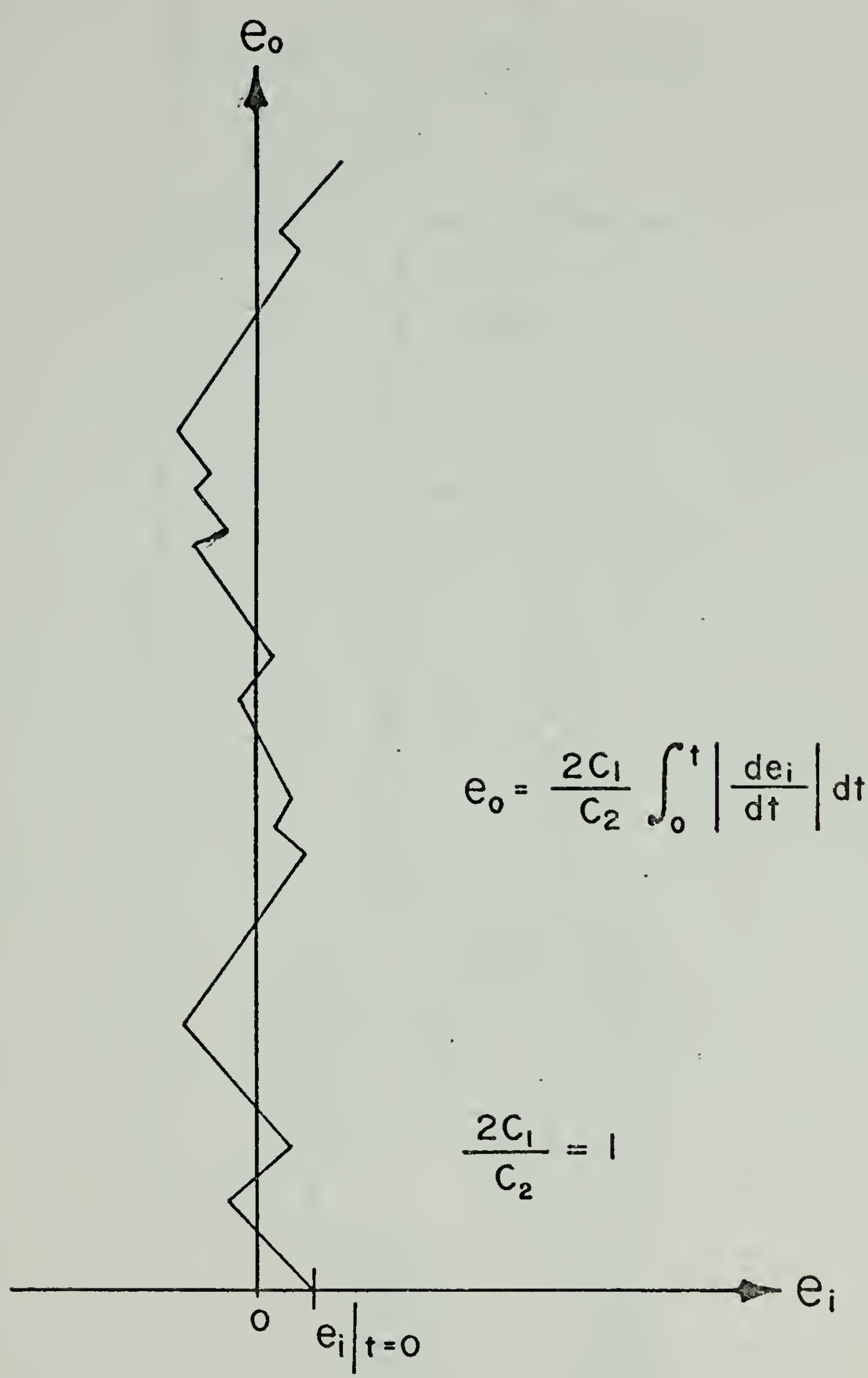


Figure 27. Signal Variation Accumulator Circuit Input-Output Characteristics.

6

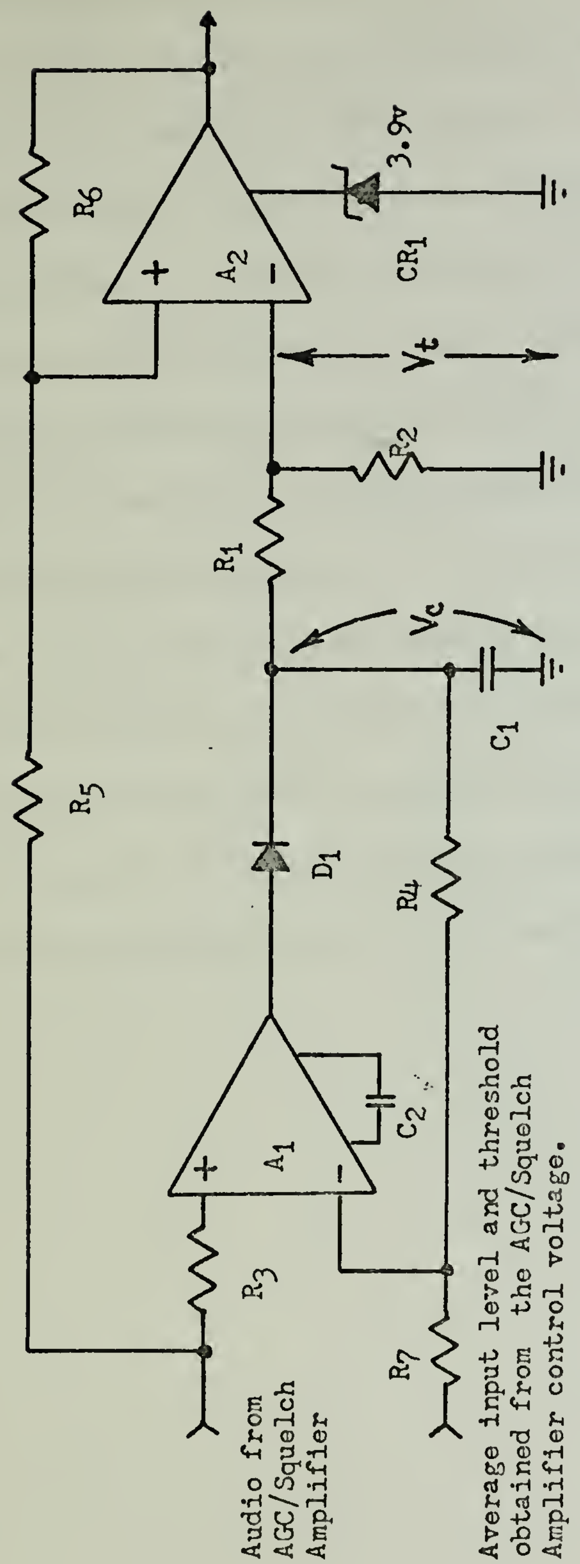


Figure 28. Adaptive Pitch Extractor Ciccuit.

Operational amplifier A_1 acts as a peak detector. When an acoustic signal is applied at the input, the amplifier charges C_1 until the voltage across it (V_c) equals the peak amplitude. The $(R_1 + R_2)C_1$ time constant is chosen to be several times the pitch period. Voltage divider R_1 and R_2 applies V_t, a portion of voltage V_c to comparator A_2 as a threshold voltage. Resistor R_6 provides hysteresis for clean switching while zener diode CR_1 clamps the output to a logic-compatible level.

The circuit responds to repeative peaks in the input waveform whose period approximates the expected pitch period. The average input level and control voltage obtained from the AGC/Squelch Amplifier is summed with V_c and applied to A_1 to provide a reference level for the peak detector. The circuit has exhibited satisfactory results for the applications described in this thesis while being relatively simple and economical. It is not recommended for highly accurate pitch determination.

V. CONCLUSIONS

A stated objective of this thesis was to develop a small, low cost speech waveform display/analyzer to be used primarily for speech aids for the handicapped. The estimated cost of the basic display as implemented is less than $150.00, exclusive of the storage CRT. The system has not yet been subjected to clinical evaluation in the training of deaf children to vocalize. However, several simple experiments with several children ranging in age from 5 years to 10 years who have normal hearing have been conducted according to the following procedure. This experimenter's phonomic pattern was reduced to a colored transparent overlay which indicated the areas of the display where activity was present. These transparencies were produced by colored pens on clear film and were then overlayed over the storage CRT face. Several children were then instructed to attempt to produce an approximate match to the overlay by saying the phonemes. (Only unvoiced speech phonemes were investigated.) The children were able to obtain visually acceptable matches within a few attempts. The perceptual quality of the phonemes was compared before and after training. This feedback technique appeared to encourage better enunciation and voice control. Total time-length of an utterance was quickly matched within ten attempts, generally.

Despite the encouraging results of these investigations, one can not generalize the results to speech training for the deaf. Much more

formal and intensive evaluation will be required. The results do appear encouraging for possible applications in foreign language training of individuals with normal hearing and articulation.

The feasibility of automatic speech recognition should be investigated using the techniques outlined in the recommendations. A limited vocabulary speech recognizer if implementable from these concepts will be small and economical enough to be made available to severely physically handicapped for control of normal household appliances (television, etc.).

Since the success of speech recognition systems will depend upon the proper selection of speech parameters to be measured, an important problem to be considered in evaluating voice parameter extractors is the comparative evaluation of alternative extraction techniques. The Time-Domain Waveform Display/Analyzer System for experimentally comparing speech time-domain parameter extractors should provide valuable results concerning the virtues of various techniques, and how they might be iteratively related or combined to yield better and better extractors.

This system is a versatile speech researcher's tool for trying out new ideas and for evaluating extractors, recognition techniques, and display feedback approaches.

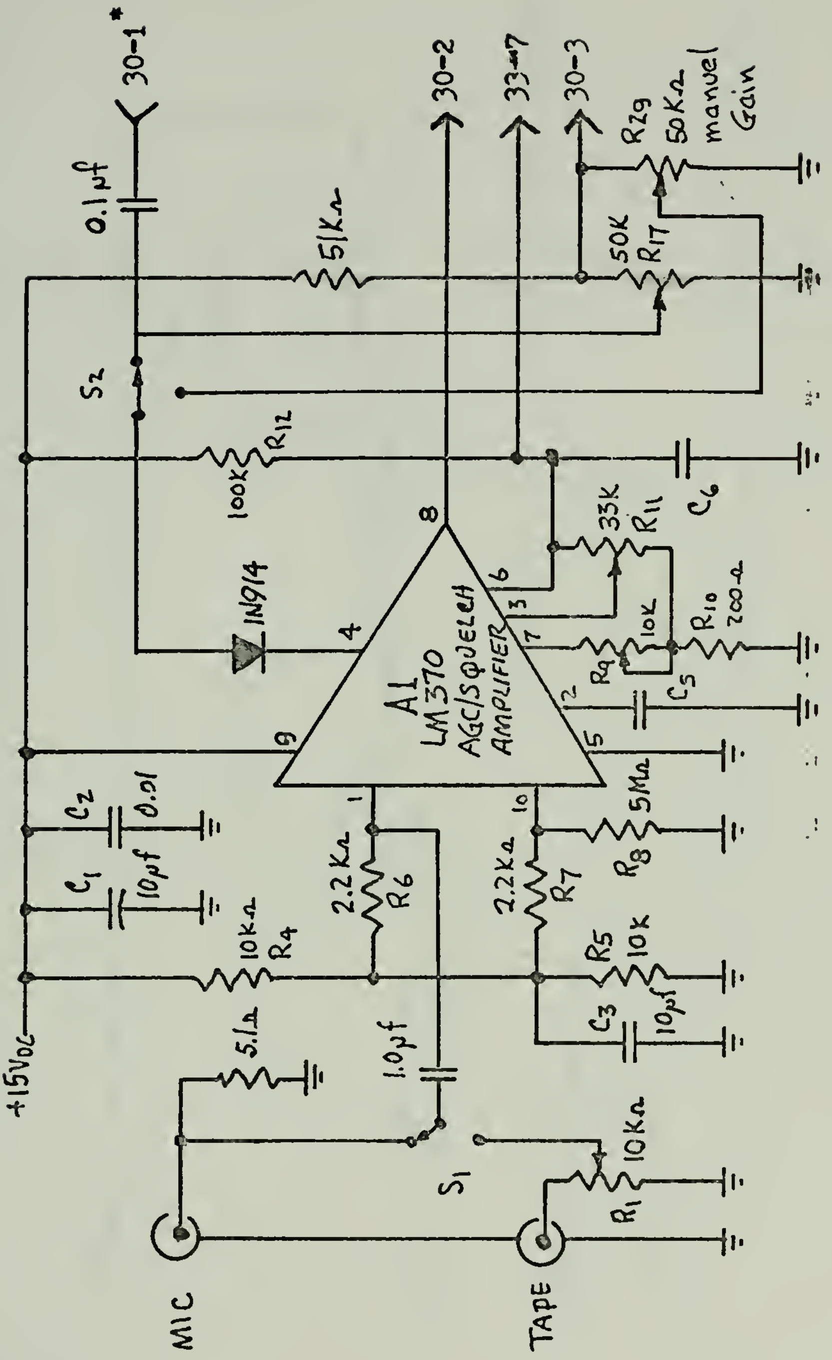
APPENDIX

A. AGC/SQUELCH AMPLIFIER [41]

Since A1 is a high gain device, power supply decoupling is provided by C1, C2, and C3 (Figure 29) while common mode input baising at one-half Vcc is provided by R4, R5, R6 and R7. The high-level input gain control, R1, reduces the input to A1 to below 50 millivolts to prevent distortion. A controlled DC offset is introduced by R8 to prevent a DC shift in the output, associated with the change in applied AGC control voltage, from driving the following gain stages into limiting causing extreme distortion.

Figure 30 shows a diagram of the AGC control loop. Amplifier A2 and A3 increase the forward gain of the AGC loop by 26 Db resulting in tighter output regulation, and buffers the AGC feedback loop from the following band-pass filter. There are two modes of operation, manual gain control, and automatic gain control, selected by S2. Connected in the AGC mode as shown in Figure 29, an emitter follower at pin 4 of A1 is used as a high impedence peak detector, with detector smoothing performed by C5 at pin 2. DC threshold for the detector is preset at one volt by potententiometer R17, determining the positive peak output voltage that initiates gain reduction. Potentiometer R29 controls amplifier gain in the manual mode.

When the input signal exceeds the threshold set by the VOX threshold potentiometer, R9, pin 6 of A1 is shorted to ground through an internal high current transistor assuring rapid discharging of squelch capacitor C6. The



*Note: Input/output pin numbers are figure number (first two digits) and corresponding pin number on that figure (after dash number). This is true for all drawings in Appendix.

Figure 29. AGC/Squelch Amplifier Schematic

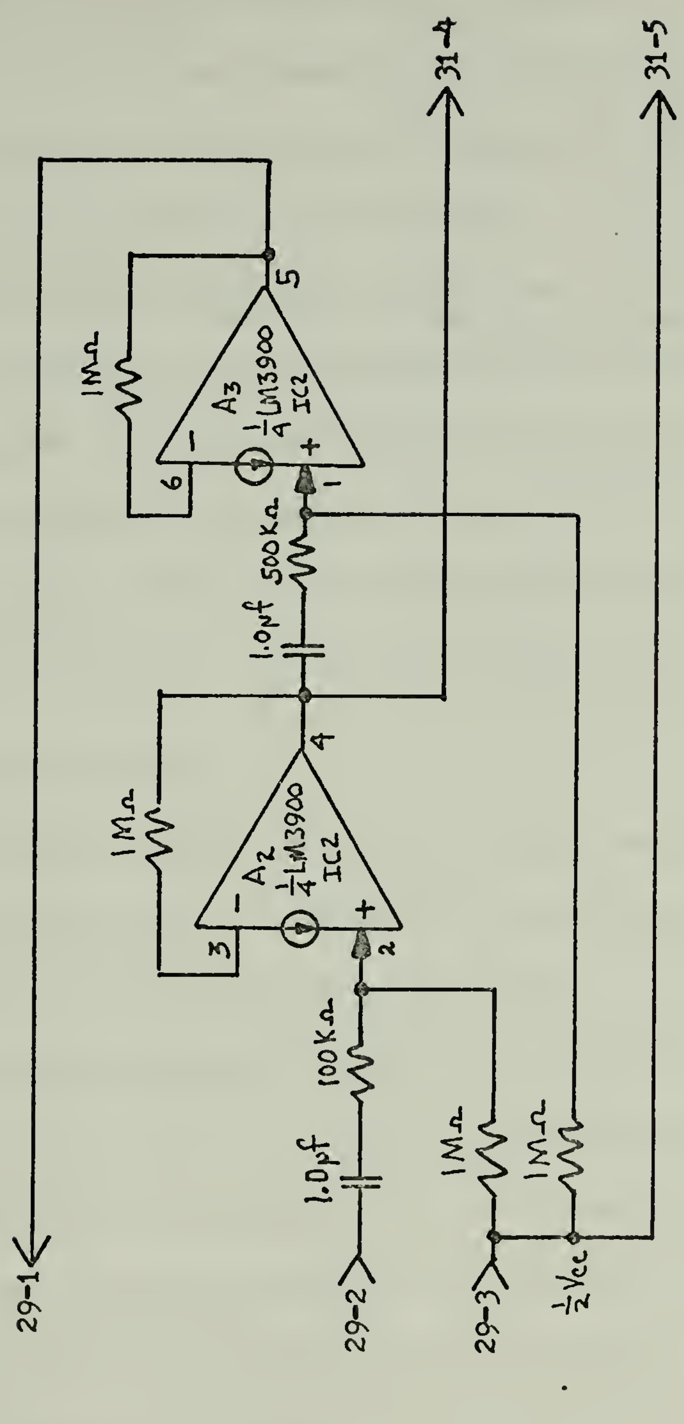


Figure 30. AGC Loop Amplifier Schematic

voltage at pin 6 is used as the Voice Operated Switch (VOX) input to the Timing Control Generator. A controlled amount of hysteresis is introduced by R10 which greatly enhances the circuits immunity to erratic speech patterns. While the Signal present signal exhibits abrupt turn-on, turn-off follows the slow charging contour of the R12-C6 time constant until the TTL logic voltage threshold is exceeded.

Under no-signal conditions and a DC threshold of one volt, the gain increases to above 40 Db and undesirably increases low level background noise during periods of silence that exceed the release time of the AGC circuit.Potentiometer, R11 reduces the no signal gain of amplifier A1 to a value, that in conjunction with the crossing detector threshold voltage, provides the degree of low level noise immunity desired.

B. BANDPASS FILTER [42]

Amplifier A4 in Figure 31 is connected as a two-pole active low-pass filter. The resistor, R4, is used to set the output bias level and is selected after the other resistors have been established.

The design procedure for this filter is to select the pass-band gain, Ho, the Q, and the corner frequency, f_c. A Q value of 1 gives only a slight peaking near the bandage ($<$ 2 db) and smaller Q values decrease this peaking. The slope of the skirt of this filter is 12 db/octive (or 40 db/ decade). The design procedure is as follows:

Given: H_o, Q, and $\omega_c = 2\pi f_c$

To find: R1, R2, R3, R4, C1, and C2

Let C1 be a convenient value, then

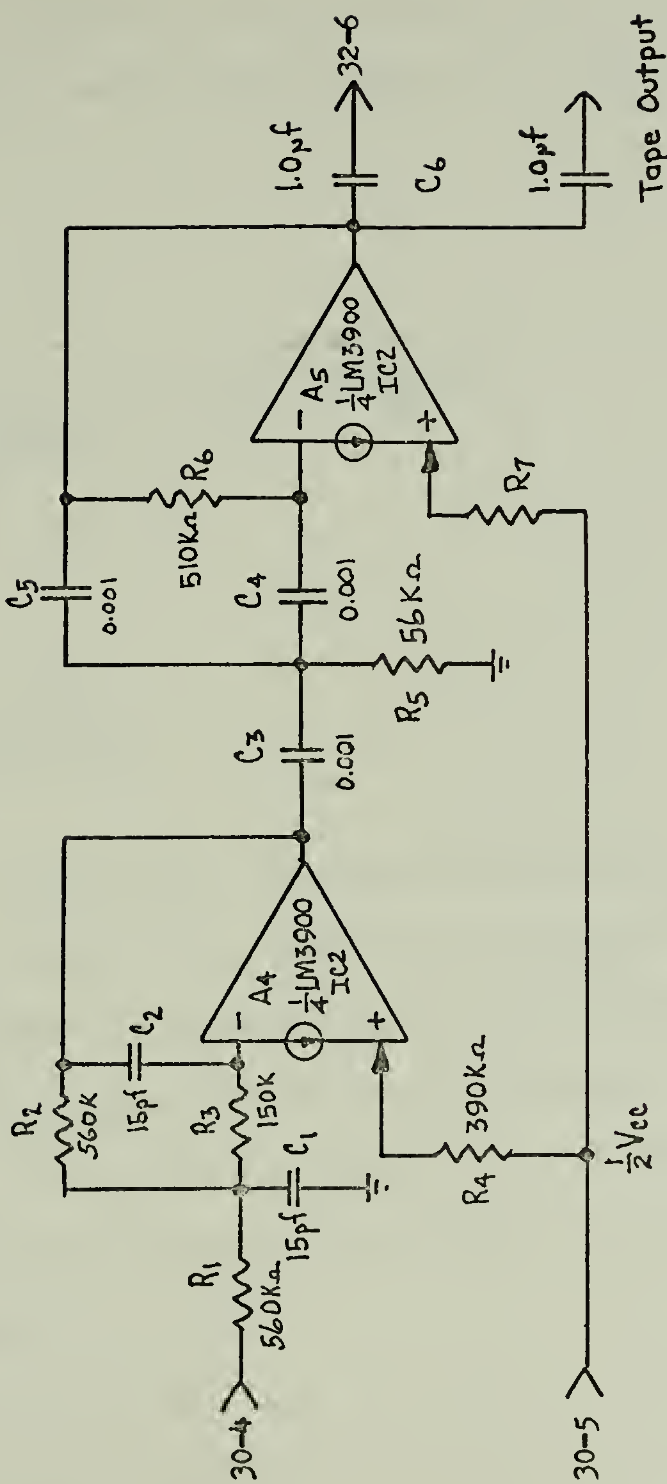


Figure 31. Band-pass Filter Schematic

$$C2 = KC1 \tag{A-1}$$

where K is a constant which can be used to adjust component values. Large values of K can be used to reduce R2 and R3 at the expense of a larger value for C2.

$$R1 = \frac{R2}{H_o} \tag{A-2}$$

$$R2 = \frac{1}{2Q\omega_c C_1}\left[1 \pm \sqrt{1 + \frac{4Q^2(H_o+1)}{K}}\right] \tag{A-3}$$

and

$$R3 = \frac{1}{\omega_c^2 C_1^2 R_2 K} \tag{A-4}$$

Amplifier A5 in Figure 31 is connected as a two-pole high-pass active filter. The resistor, R7, is made equal to R6 and a Bias reference of $1/2V_{CC}$ will establish the output Q point at this value ($1/2V_{CC}$). The input is capacitively coupled (C3) and there are therefore no further DC biasing problems. The slope of the skirt of this filter is 12 db/octive. If the gain H_o is unity all capacitors have the same value. The design proceeds as follows:

Given: H_o, Q and $\omega_c = 2\pi f_c$

To find: R5, R6, R7, C3, C4, and C5

Let C1 = C3 and choose a convenient starting value. R7 = R6.

Then:

$$R1 = \frac{1}{Q\omega_c C_1 (2H_o + 1)} \tag{A-5}$$

$$R2 = \frac{Q}{\omega_c C_1}(2H_o + 1) \tag{A-6}$$

and

$$C2 = \frac{C_1}{H_o} \tag{A-7}$$

C. LOGARITHMIC THRESHOLD CROSSING DETECTOR

The high frequency response and good DC characteristics of a CA3018A monolithic transistor array were used to design a fast logarithmic-feedback threshold-crossing detector that works over a 50 Db range in input variation. The array, connected as an operational amplifier, has a TTL-compatible output.

Transistors Q1 and Q2 in Figure 32 are a differential input pair, with Q2 loaded by R6. Transistor Q3 acts as an output emitter follower. Resistor R6 is boot-strapped to zener-connector Q4, enabling the single stage of amplification to provide more than 50 db of voltage gain.

Negative feedback through diodes CR1-CR4 constrains the peak output limits to twice the forward voltage drop per diode about the TTL transition region of 1.4 V. The negative feedback and the logarithmic variation of the output voltage with increasing signal input voltage allow a wide range of input variation. Positive feedback through resistor R5 causes the circuit to have a defined lower threshold and a snap-action transition, allowing noise-free threshold crossing detection near zero volts. R2 is the threshold adjustment which offsets the voltage center of the input transition by varying the bias on Q2.

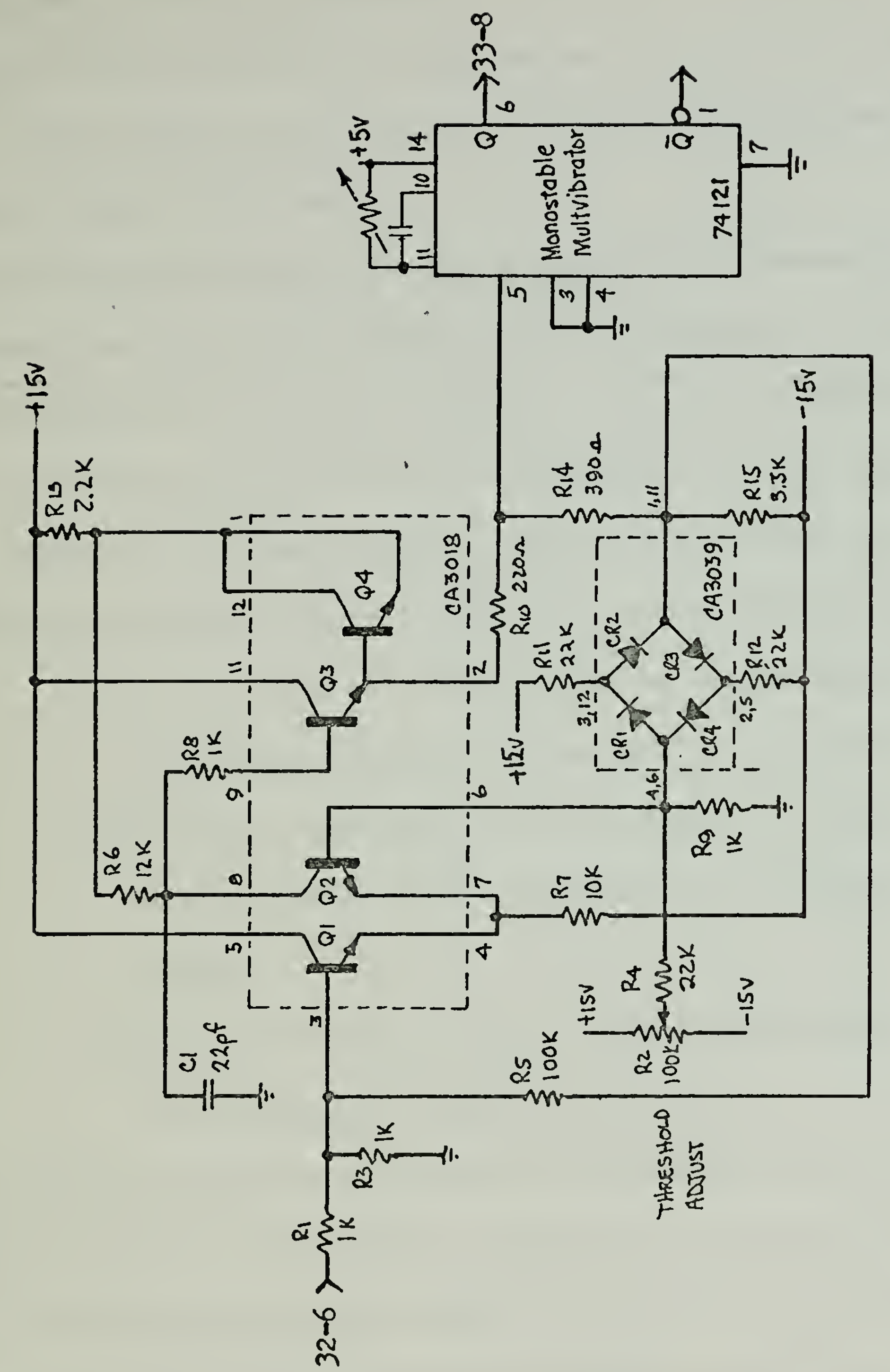


Figure 32. Logarithmic Threshold
Crossing Detector Schematic

D. TIMING CONTROL GENERATOR

The timing function of this circuit may best be understood by referring to the simplified block diagram, Figure 33. The blocks labeled MV (.) are TTL monostable multivibrators.

Multivibrator MV1 provides a constant preset delay between the up-crossing transition and the beginning of parameterization. This delay may be preset from 0-70 μsec to allow the up-crossing transition to be used as a trigger input to follow-on systems (described in the recommendation section of this thesis) before the parameterization process is begun.

The retriggerable multivibrator MV2 provides the following functions:

(1) In conjunction with MV2 inhibits the Display Strobe for up-crossing intervals less than 100 μs in duration;

(2) Provides an integrator reset and a begin conversion signal to the Up-Crossing Interval to Log-Frequency Converter;

(3) Together with the EOC signal, gates a 2 MHz clock to the Vertical Display Counter-D/A Converter.

A fixed time delay of 70 μs after the delayed up-crossing transition is generated at the Q output of MV2. The first 64 μs is used for parameterization of the up-crossing interval while the remaining 6 μs allow the Vertical Display Counter-D/A converter to settle to its final value. If the VOX signal is ON (signal present), the positive transition of the Q output of MV2 triggers MV4 to produce a 10 μs display strobe and horizontal

Display Counter clock pulse. The Q output of MV4 triggers MV5 to produce a 1 μs Vertical Display Counter reset pulse. MV6 and MV7 form a gated 2Mhz clock.

The output drive pulses are level shifted to drive CMOS logic by open-collector inverters I1-I6. The NE556 oscillator is used to drive the Horizontal Display Counter in the real-time mode.

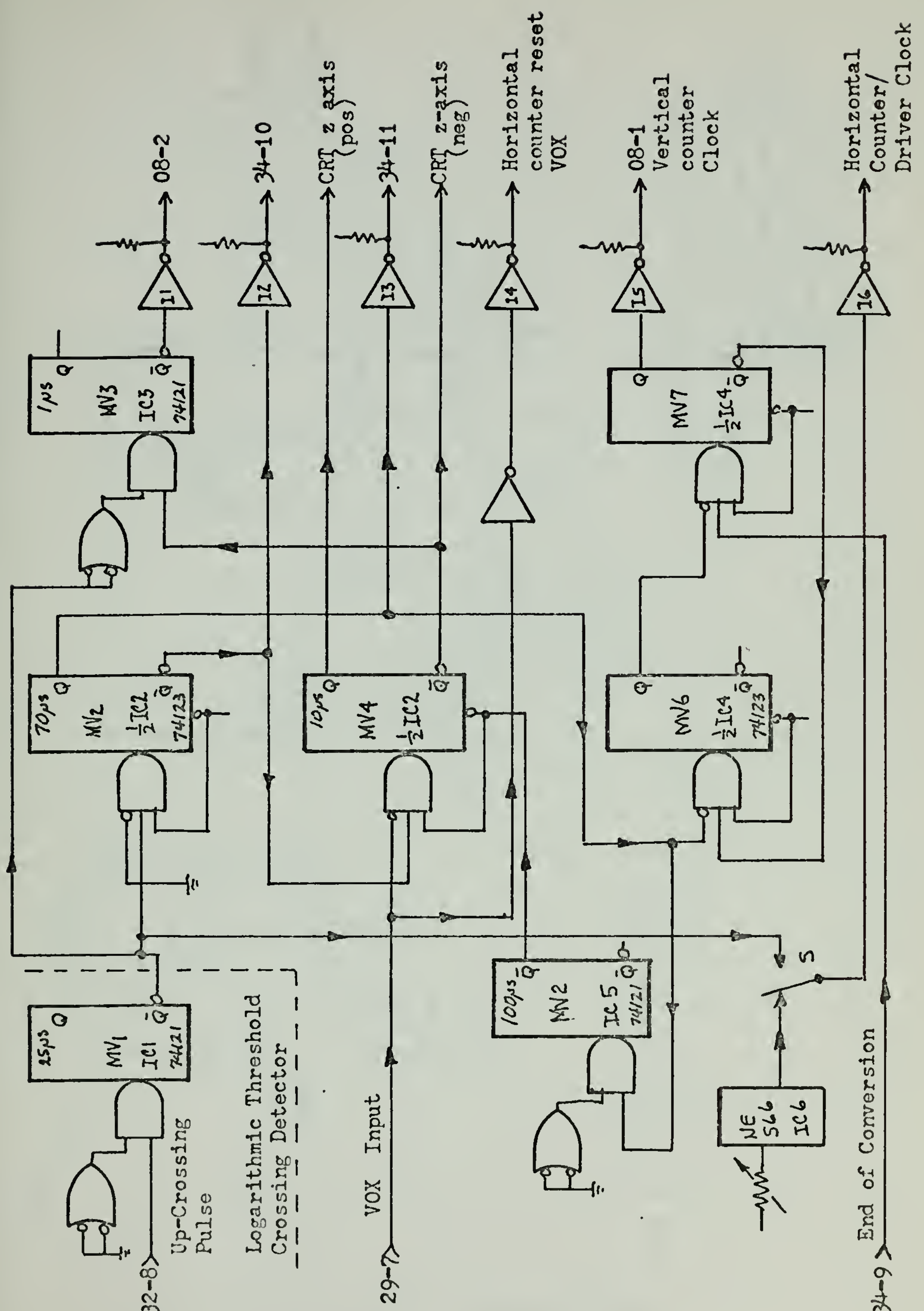


Figure 33. Timing Control Generator.

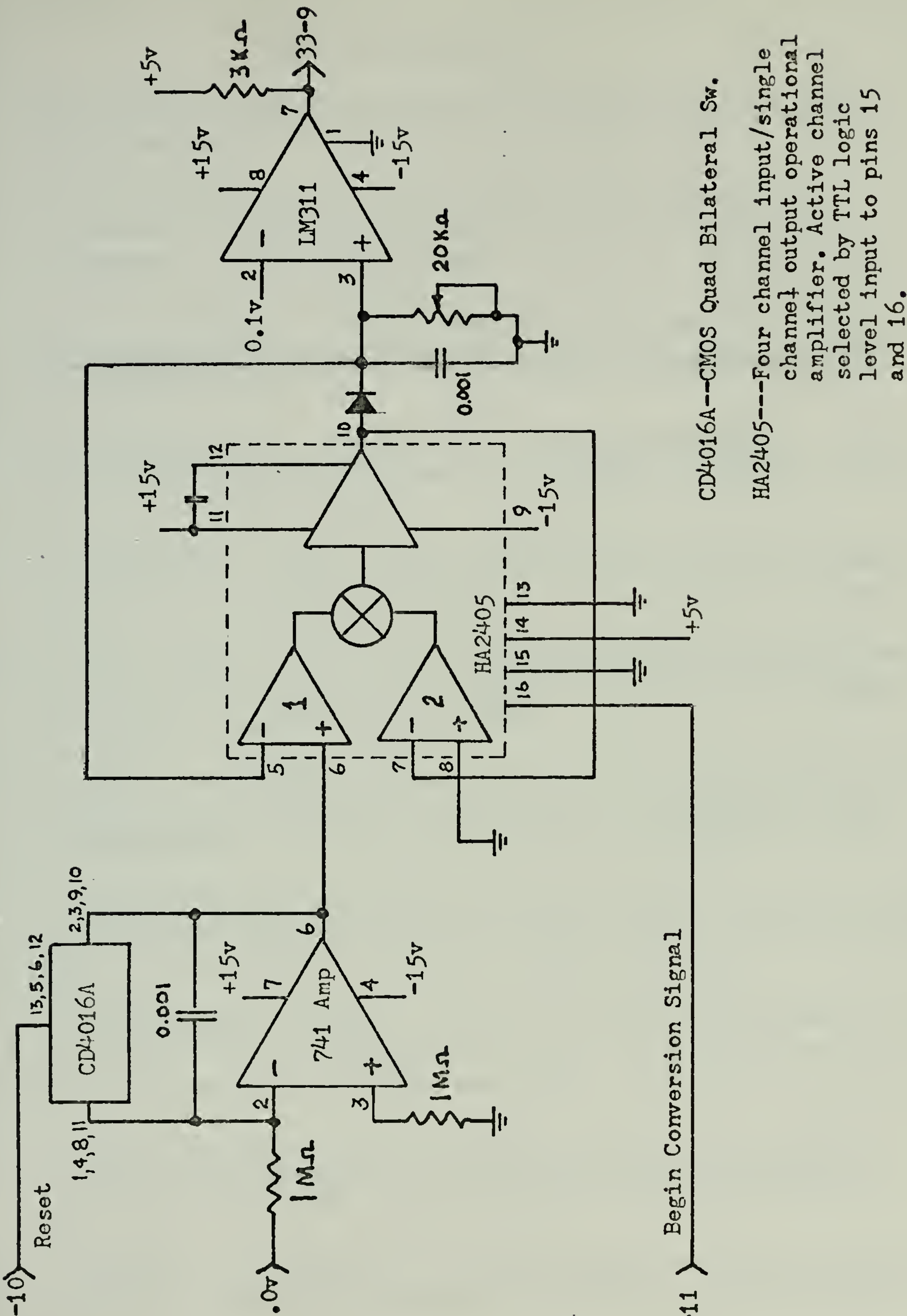


Figure 34. Log RCI Schematic Diagram

LIST OF REFERENCES

1. N. Lindgren, "Machine Recognition of Human Language, Part I," IEEE Spectrum, vol. 24 , p. 114-136, March 1965.

2. J.C.R. Licklider and I. Pollack, "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," J. Acoust. Soc. Amer., vol. 20, p. 42-51, 1948.

3. J.C.R. Licklider, "The Intelligibility of Amplitude-dichotomized Time-Quantized Speech Wave," J. Acoust. Soc. Amer., vol. 22, p. 820-823, 1950.

4. S. Chang, G. Pihl, and J. Wiren, "The Intervalgram as a Visual Representation of Speech Sounds," J. Acoust. Soc. Amer., vol. 23, no. 6, p. 675-679, 1951.

5. T. Sakai and S. Inoue, "New Instruments and Methods for Speech Analysis, "J. Acoust. Soc. Amer., vol. 32, p. 441-450, 1960.

6. R.W.A. Scarr, "Zero Crossings as a Means of Obtaining Spectral Information in Speech Analysis," IEEE Trans. Audio Electroacoust., vol. AU-16, p. 247-255, June 1968.

7. M. R. Ito and R. W. Donaldson, "Zero-Crossing Measurements for Analysis and Recognition of Speech Sounds," IEEE Trans. Audio Electroacoust., vol. AU-19, p. 235-242, Sept. 1971.

8. W. Bezdel and B. A. Chandler, "Results of an Analysis and Recognition of Vowels by Computer Using Zero-Crossing Data," Proc. Inst. Elect. Eng., vol. 112, no. 11, p. 2060-2066, Nov. 1965.

9. W. Bezdel and J. S. Bridle, "Speech Recognition Using Zero-Crossing Measurements and Sequence Information," Proc. Inst. Elect. Eng., vol. 116, no. 4, p. 617-623, April 1969.

10. G. D. Ewing and J. F. Taylor, "Computer Recognition of Speech using Zero-Crossing Information," IEEE Trans. Audio Electroacoust., vol. AU-17, p. 37-40, March 1969.

11. R. De Mori, "A Descriptive Technique for Automatic Speech Recognition," IEEE Trans. Audio Electroacoust., vol. AU-21, p. 89-100, April 1973.

12. D. R. Reddy, "Segmentation of Speech Sounds," J. Acoust. Soc. Amer., vol. 40, p. 307-312, 1966.

13. W. Koenig, H. K. Dunn, and L. Y. Lacy, "The Sound Spectrograph," J. Acoust. Soc. Amer., vol. 18, p. 19-49, 1946.

14. W. H. Huggins, "A Phase Principle for Complex-Frequency Analysis," J. Acoust. Soc. Amer., vol. 24, p. 582-589, 1952.

15. W. E. Kock and R. L. Miller, "Dynamic Spectrograms of Speech," J. Acoust. Soc. Amer., vol. 24, p. 783-784, 1952.

16. H. Dudley and O. Gruenz, Jr., "Visible Speech Translators with External Phosphors," J. Acoust. Soc. Amer., vol. 18, p. 62-73, 1946.

17. R. R. Riesz and L. Scott, "Visible Speech Cathode-ray Translator," J. Acoust. Soc. Amer., vol. 18, p. 50-61, 1946.

18. W. R. Bennet, "The Correlatograph," Bell Systems Tech. Jour., vol. 32, p. 1173-1185, 1953.

19. L. G. Kersta, "Amplitude Cross-section Representation with the Sound Spectrograph," J. Acoust. Soc. Amer., vol. 20, p. 796-801, 1948.

20. A. J. Prestigiacomo, "Amplitude Controu Display of Sound Spectrograms," J. Acoust. Soc. Amer., vol. 34, p. 1684-1688, 1962.

21. J. Baker, J. M. Baker, and J. Littvin, "More Visible Speech," J. Acoust. Soc. Amer., vol. 52, no. 2, part 1, p. 183, July 1972.

22. S. Cantarano and G. V. Pallottino, "Logarithmic Analog-to-Digital Converters: A Survey," IEEE Trans. Instrum. Meas., vol. IM-22, p. 201-213, Sept. 1973.

23. J. F. Gobbons and H. S. Horn, "A Circuit with Logarithmic Transfer Response over 9 Decades," IEEE Trans. Circuit Theory, vol. CT-11, p. 378-384, Sept. 1964.

24. R. C. Dobkins, "Logarithmic Converters," IEEE Spectrum, vol. 6, p. 69-72, Nov. 1969.

25. E. M. Glaser and H. Blasbalg, "A Logarithmic Voltage Quantizer," IRE Trans. Electron. Comput., vol. EC-4, p. 150-155, Dec. 1955.

26. E. J. Duke, "RC Logarithmic Analog-to-Digital (LAD) Conversion," IEEE Trans. Instrum. Meas., vol. IM-20, p. 74-76, Feb. 1971.

27. D. Degryse and B. Guerin, "A Logarithmic Transcoder," IEEE Trans. Comput., vol. C-21, p. 1165-1168, Nov. 1972.

28. Radio Corporation of America, COS/MOS Integrated Circuits Manual, CMS-271, p. 125, RCA Solid State Division, June 1972.

29. M. Halle and K. N. Stevens, "Speech Recognition: A Model and a Program for Research," in The Structure of Language, J. A. Fador and J. J. Katz, Eds., p. 604-612, Prentice-Hall, 1964.

30. N. Nilsson, Learning Machines, New York: McGraw-Hill, 1965.

31. Ibid., p. 4.

32. R. K. Potter, G. A. Kopp, and H. C. Green, Visible Speech, New York: D. van Nostrand Co., 1947.

33. J. L. Steward, Frequency Filtering Processes in the Human Ear, published by Santa Rita Technology, Menlo Park, Calif., Jan. 1966.

34. L. R. Focht, The Single Equivalent Formant, Digest of Technical Papers, IEEE International Communication Conference, June 1966.

35. C. F. Teacher, H. G. Kellett, and L. R. Focht, "Experimental Limited Vocabulary, Speech Recognizer," IEEE Trans. Audio Electroacoust., vol. AU-15, p. 127-130, Sept. 1967.

36. W. C. Dersch, "Shoebox: A Voice Responsive Machine," Datamation, June 1962.

37. M. M. Sondihi, "New Methods of Pitch Extraction," IEEE Trans. Audio Electroacoust., vol. AU-16, p. 262-266, June 1968.

38. M. R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement," J. Acoust. Soc. Amer., vol. 43, no. 4, 1968.

39. A. M. Noll, "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-pitch Detection," J. Acoust. Soc. Amer., vol. 36, Feb. 1964.

40. J. N. Maksym, "Real-Time Pitch Extraction by Adaptive Prediction of the Speech Waveform," IEEE Trans. Audio Electroacoust., vol. 21, p. 149-154, June 1973.

41. National Semiconductor Linear Applications Note No. 51, A Unique Monolithic AGC/Squelch Amplifier, by T. Smathers, Sept. 1971.

42. National Semiconductor Linear Applications Note No. 72, The LM3900: A New Current-Differencing Quad of + Input Amplifiers, T. M. Frederiksen, W. M. Howard, and R. S. Sleeth, Sept. 1972.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Department Chairman, Code 52 Department of Electrical Engineering Naval Postgraduate School Monterey, California 93940	2
4.	Asst Professor G. D. Ewing, Code 52Ew Department of Electrical Engineering Naval Postgraduate School Monterey, California 93940	1
5.	LT Gary Duane Edmondson 1505 Pepper Drive El Cajon, California 92022	1